# 1
# Fundamental concepts

## 1.1 Probability and random variables

The aim of this book is to present the most important concepts and methods of statistical data analysis. A central concept is that of uncertainty, which can manifest itself in a number of different ways. For example, one is often faced with a situation where the outcome of a measurement varies unpredictably upon repetition of the experiment. Such behavior can result from errors related to the measuring device, or it could be the consequence of a more fundamental (e.g. quantum mechanical) unpredictability of the system. The uncertainty might stem from various undetermined factors which in principle could be known but in fact are not. A characteristic of a system is said to be **random** when it is not known or cannot be predicted with complete certainty.

The degree of randomness can be quantified with the concept of **probability**. The mathematical theory of probability has a history dating back at least to the 17th century, and several different definitions of probability have been developed. We will use the definition in terms of set theory as formulated in 1933 by Kolmogorov [Kol33]. Consider a set $S$ called the **sample space** consisting of a certain number of elements, the interpretation of which is left open for the moment. To each subset $A$ of $S$ one assigns a real number $P(A)$ called a probability, defined by the following three axioms:[1]

(1) For every subset $A$ in $S$, $P(A) \geq 0$.

(2) For any two subsets $A$ and $B$ that are disjoint (i.e. mutually exclusive, $A \cap B = \emptyset$) the probability assigned to the union of $A$ and $B$ is the sum of the two corresponding probabilities, $P(A \cup B) = P(A) + P(B)$.

(3) The probability assigned to the sample space is one, $P(S) = 1$.

From these axioms further properties of probability functions can be derived, e.g.

[1] The axioms here are somewhat simplified with respect to those found in more rigorous texts, such as [Gri92], but are sufficient for our purposes. More precisely, the set of subsets to which probabilities are assigned must constitute a so-called $\sigma$-field.

$$\begin{array}{l}P(\overline{A}) = 1 - P(A) \text{ where } \overline{A} \text{ is the complement of } A \\ P(A \cup \overline{A}) = 1 \\ 0 \leq P(A) \leq 1 \\ P(\emptyset) = 0 \\ \text{if } A \subset B, \text{ then } P(A) \leq P(B) \\ P(A \cup B) = P(A) + P(B) - P(A \cap B).\end{array} \tag{1.1}$$

For proofs and further properties see e.g. [Bra92, Gri86, Gri92].

A variable that takes on a specific value for each element of the set $S$ is called a **random variable**. The individual elements may each be characterized by several quantities, in which case the random variable is a multicomponent vector.

Suppose one has a sample space $S$ which contains subsets $A$ and $B$. Provided $P(B) \neq 0$, one defines the **conditional probability** $P(A|B)$ (read $P$ of $A$ given $B$) as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.2}$$

Figure 1.1 shows the relationship between the sets $A$, $B$ and $S$. One can easily show that conditional probabilities themselves satisfy the axioms of probability. Note that the usual probability $P(A)$ can be regarded as the conditional probability for $A$ given $S$: $P(A) = P(A|S)$.

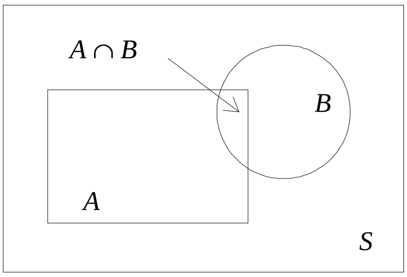

**Fig. 1.1** Relationship between the sets $A$, $B$ and $S$ in the definition of conditional probability.

Two subsets $A$ and $B$ are said to be **independent** if

$$P(A \cap B) = P(A)\,P(B). \tag{1.3}$$

For $A$ and $B$ independent, it follows from the definition of conditional probability that $P(A|B) = P(A)$ and $P(B|A) = P(B)$. (Do not confuse independent subsets according to (1.3) with disjoint subsets, i.e. $A \cap B = \emptyset$.)

From the definition of conditional probability one also has the probability of $B$ given $A$ (assuming $P(A) \neq 0$),

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{1.4}$$

Since $A \cap B$ is the same as $B \cap A$, by combining equations (1.2) and (1.4) one has

$$P(B \cap A) = P(A|B)\,P(B) = P(B|A)\,P(A), \tag{1.5}$$

or

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}. \tag{1.6}$$

Equation (1.6), which relates the conditional probabilities $P(A|B)$ and $P(B|A)$, is called **Bayes' theorem** [Bay63].

Suppose the sample space $S$ can be broken into disjoint subsets $A_i$, i.e. $S = \cup_i A_i$ with $A_i \cap A_j = \emptyset$ for $i \neq j$. Assume further that $P(A_i) \neq 0$ for all $i$. An arbitrary subset $B$ can be expressed as $B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$. Since the subsets $B \cap A_i$ are disjoint, their probabilities add, giving

$$\begin{aligned} P(B) &= P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i) \\ &= \sum_i P(B|A_i) P(A_i). \end{aligned} \tag{1.7}$$

The last line comes from the definition (1.4) for the case $A = A_i$. Equation (1.7) is called the **law of total probability**. It is useful, for example, if one can break the sample space into subsets $A_i$ for which the probabilities are easy to calculate. It is often combined with Bayes' theorem (1.6) to give

$$P(A|B) = \frac{P(B|A)\,P(A)}{\sum_i P(B|A_i) P(A_i)}. \tag{1.8}$$

Here $A$ can be any subset of $S$, including, for example, one of the $A_i$.

As an example, consider a disease which is known to be carried by 0.1% of the population, i.e. the **prior probabilities** to have the disease or not are

$$\begin{aligned} P(\text{disease}) &= 0.001, \\ P(\text{no disease}) &= 0.999. \end{aligned}$$

A test is developed which yields a positive result with a probability of 98% given that the person carries the disease, i.e.

$$\begin{aligned} P(+|\text{disease}) &= 0.98, \\ P(-|\text{disease}) &= 0.02. \end{aligned}$$

Suppose there is also a 3% probability, however, to obtain a positive result for a person without the disease,

$$
\begin{aligned}
P(+|\text{no disease}) &= 0.03,\\
P(-|\text{no disease}) &= 0.97.
\end{aligned}
$$

What is the probability that you have the disease if your test result is positive? According to Bayes' theorem (in the form of equation (1.8)) this is given by

$$
\begin{aligned}
P(\text{disease}|+) &= \frac{P(+|\text{disease})\,P(\text{disease})}{P(+|\text{disease})\,P(\text{disease})\;+\;P(+|\text{no disease})\,P(\text{no disease})}\\
&= \frac{0.98\times 0.001}{0.98\times 0.001+0.03\times 0.999}\\
&= 0.032.
\end{aligned}
$$

The probability that you have the disease given a positive test result is only 3.2%. This may be surprising, since the probability of having a wrong result is only 2% if you carry the disease and 3% if you do not. But the prior probability is very low, 0.1%, which leads to a **posterior probability** of only 3.2%. An important point that we have skipped over is what it means when we say $P(\text{disease}|+) = 0.032$, i.e. how exactly the probability should be interpreted. This question is examined in the next section.

## 1.2 Interpretation of probability

Although any function satisfying the axioms above can be called by definition a probability function, one must still specify how to interpret the elements of the sample space and how to assign and interpret the probability values. There are two main interpretations of probability commonly used in data analysis. The most important is that of **relative frequency**, used among other things for assigning statistical errors to measurements. Another interpretation called **subjective** probability is also used, e.g. to quantify systematic uncertainties. These two interpretations are described in more detail below.

### 1.2.1 Probability as a relative frequency

In data analysis, probability is most commonly interpreted as a **limiting relative frequency**. Here the elements of the set $S$ correspond to the possible outcomes of a measurement, assumed to be (at least hypothetically) repeatable. A subset $A$ of $S$ corresponds to the occurrence of any of the outcomes in the subset. Such a subset is called an **event**, which is said to occur if the outcome of a measurement is in the subset.

A subset of $S$ consisting of only one element denotes a single **elementary outcome**. One assigns for the probability of an elementary outcome $A$ the fraction of times that $A$ occurs in the limit that the measurement is repeated an infinite number of times:

$$P(A) = \lim_{n \to \infty} \frac{\text{number of occurrences of outcome } A \text{ in } n \text{ measurements}}{n}. \quad (1.9)$$

The probabilities for the occurrence of any one of several outcomes (i.e. for a non-elementary subset $A$) are determined from those for individual outcomes by the addition rule given in the axioms of probability. These correspond in turn to relative frequencies of occurrence.

The relative frequency interpretation is consistent with the axioms of probability, since the fraction of occurrences is always greater than or equal to zero, the frequency of any out of a disjoint set of outcomes is the sum of the individual frequencies, and the measurement must by definition yield some outcome (i.e. $P(S) = 1$). The conditional probability $P(A|B)$ is thus the number of cases where both $A$ and $B$ occur divided by the number of cases in which $B$ occurs, regardless of whether $A$ occurs. That is, $P(A|B)$ gives the frequency of $A$ with the subset $B$ taken as the sample space.

Clearly the probabilities based on such a model can never be determined experimentally with perfect precision. The basic tasks of **classical statistics** are to estimate the probabilities (assumed to have some definite but unknown values) given a finite amount of experimental data, and to test to what extent a particular model or theory that predicts probabilities is compatible with the observed data.

The relative frequency interpretation is straightforward when studying physical laws, which are assumed to act the same way in repeated experiments. The validity of the assigned probability values can be experimentally tested. This point of view is appropriate, for example, in particle physics, where repeated collisions of particles constitute repetitions of an experiment. The concept of relative frequency is more problematic for unique phenomena such as the big bang. Here one can attempt to rescue the frequency interpretation by imagining a large number of similar universes, in some fraction of which a certain event occurs. Since, however, this is not even in principle realizable, the frequency here must be considered as a mental construct to assist in expressing a degree of belief about the single universe in which we live.

The frequency interpretation is the approach usually taken in standard texts on probability and statistics, such as those of Fisher [Fis90], Stuart and Ord [Stu91] and Cramér [Cra46]. The philosophy of probability as a frequency is discussed in the books by von Mises [Mis51, Mis64].

### 1.2.2 Subjective probability

Another probability interpretation is that of **subjective** (also called **Bayesian**) probability. Here the elements of the sample space correspond to **hypotheses** or **propositions**, i.e. statements that are either true or false. (When using subjective probability the sample space is often called the hypothesis space.) One interprets the probability associated with a hypothesis as a measure of degree of belief:

$$P(A) = \text{degree of belief that hypothesis } A \text{ is true.} \quad (1.10)$$

The sample space $S$ must be constructed such that the elementary hypotheses are mutually exclusive, i.e. only one of them is true. A subset consisting of more than one hypothesis is true if any of the hypotheses in the subset is true. That is, the union of sets corresponds to the Boolean OR operation and the intersection corresponds to AND. One of the hypotheses must necessarily be true, i.e. $P(S) = 1$.

The statement that a measurement will yield a given outcome a certain fraction of the time can be regarded as a hypothesis, so the framework of subjective probability includes the relative frequency interpretation. In addition, however, subjective probability can be associated with, for example, the value of an unknown constant; this reflects one's confidence that its value lies in a certain fixed interval. A probability for an unknown constant is not meaningful with the frequency interpretation, since if we repeat an experiment depending on a physical parameter whose exact value is not certain (e.g. the mass of the electron), then its value is either never or always in a given fixed interval. The corresponding probability would be either zero or one, but we do not know which. With subjective probability, however, a probability of 95% that the electron mass is contained in a given interval is a reflection of one's state of knowledge.

The use of subjective probability is closely related to Bayes' theorem and forms the basis of **Bayesian** (as opposed to classical) statistics. The subset $A$ appearing in Bayes' theorem (equation (1.6)) can be interpreted as the hypothesis that a certain theory is true, and the subset $B$ can be the hypothesis that an experiment will yield a particular result (i.e. data). Bayes' theorem then takes on the form

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory}) \cdot P(\text{theory}).$$

Here $P(\text{theory})$ represents the **prior probability** that the theory is true, and $P(\text{data}|\text{theory})$, called the **likelihood**, is the probability, under the assumption of the theory, to observe the data which were actually obtained. The **posterior probability** that the theory is correct after seeing the result of the experiment is then given by $P(\text{theory}|\text{data})$. Here the prior probability for the data $P(\text{data})$ does not appear explicitly, and the equation is expressed as a proportionality. Bayesian statistics provides no fundamental rule for assigning the prior probability to a theory, but once this has been done, it says how one's degree of belief should change in the light of experimental data.

Consider again the probability to have a disease given a positive test result. From the standpoint of someone studying a large number of potential carriers of the disease, the probabilities in this problem can be interpreted as relative frequencies. The prior probability $P(\text{disease})$ is the overall fraction of people who carry the disease, and the posterior probability $P(\text{disease}|+)$ gives the fraction of people who are carriers out of those with a positive test result. A central problem of classical statistics is to estimate the probabilities that are assumed to describe the population as a whole by examining a finite sample of data, i.e. a subsample of the population.

A specific individual, however, may be interested in the subjective probability that he or she has the disease given a positive test result. If no other information is available, one would usually take the prior probability $P(\text{disease})$ to be equal to the overall fraction of carriers, i.e. the same as in the relative frequency interpretation. Here, however, it is taken to mean the degree of belief that one has the disease before taking the test. If other information is available, different prior probabilities could be assigned; this aspect of Bayesian statistics is necessarily subjective, as the name of the probability interpretation implies. Once $P(\text{disease})$ has been assigned, however, Bayes' theorem then tells how the probability to have the disease, i.e. the degree of belief in this hypothesis, changes in light of a positive test result.

The use of subjective probability is discussed further in Sections 6.13, 9.8 and 11.5.3. There exists a vast literature on subjective probability; of particular interest are the books by Jeffreys [Jef48], Savage [Sav72], de Finetti [Fin74] and the paper by Cox [Cox46]. Applications of Bayesian methods are discussed in the books by Lindley [Lin65], O'hagan [Oha94], Lee [Lee89] and Sivia [Siv96].

## 1.3 Probability density functions

Consider an experiment whose outcome is characterized by a single continuous variable $x$. The sample space corresponds to the set of possible values that $x$ can assume, and one can ask for the probability of observing a value within an infinitesimal interval $[x, x+dx]$.[2] This is given by the **probability density function** (p.d.f.) $f(x)$:

$$\text{probability to observe } x \text{ in the interval } [x, x+dx] = f(x)dx. \tag{1.11}$$

In the relative frequency interpretation, $f(x)dx$ gives the fraction of times that $x$ is observed in the interval $[x, x+dx]$ in the limit that the total number of observations is infinitely large. The p.d.f. $f(x)$ is normalized such that the total probability (probability of some outcome) is one,

$$\int_S f(x)dx = 1, \tag{1.12}$$

where the region of integration $S$ refers to the entire range of $x$, i.e. to the entire sample space.

Although finite data samples will be dealt with more thoroughly in Chapter 5, it is illustrative here to point out the relationship between a p.d.f. $f(x)$ and a set of $n$ observations of $x$, $x_1, \ldots, x_n$. A set of such observations can be displayed graphically as a **histogram** as shown in Fig. 1.2. The $x$ axis of the histogram is

[2] A possible confusion can arise from the notation used here, since $x$ refers both to the random variable and also to a value that can be assumed by the variable. Many authors use upper case for the random variable, and lower case for the value, i.e. one speaks of $X$ taking on a value in the interval $[x, x+dx]$. This notation is avoided here for simplicity; the distinction between variables and their values should be clear from context.

divided into $m$ subintervals or **bins** of width $\Delta x_i$, $i = 1, \ldots, m$, where $\Delta x_i$ is usually but not necessarily the same for each bin. The number of occurrences $n_i$ of $x$ in subinterval $i$, i.e. the number of entries in the bin, is given on the vertical axis. The area under the histogram is equal to the total number of entries $n$ multiplied by $\Delta x$ (or for unequal bin widths, $area = \sum_{i=1}^{m} n_i \cdot \Delta x_i$). Thus the histogram can be normalized to unit area by dividing each $n_i$ by the corresponding bin width $\Delta x_i$ and by the total number of entries in the histogram $n$. The p.d.f. $f(x)$ corresponds to a histogram of $x$ normalized to unit area in the limit of zero bin width and an infinitely large total number of entries, as illustrated in Fig. 1.2(d).

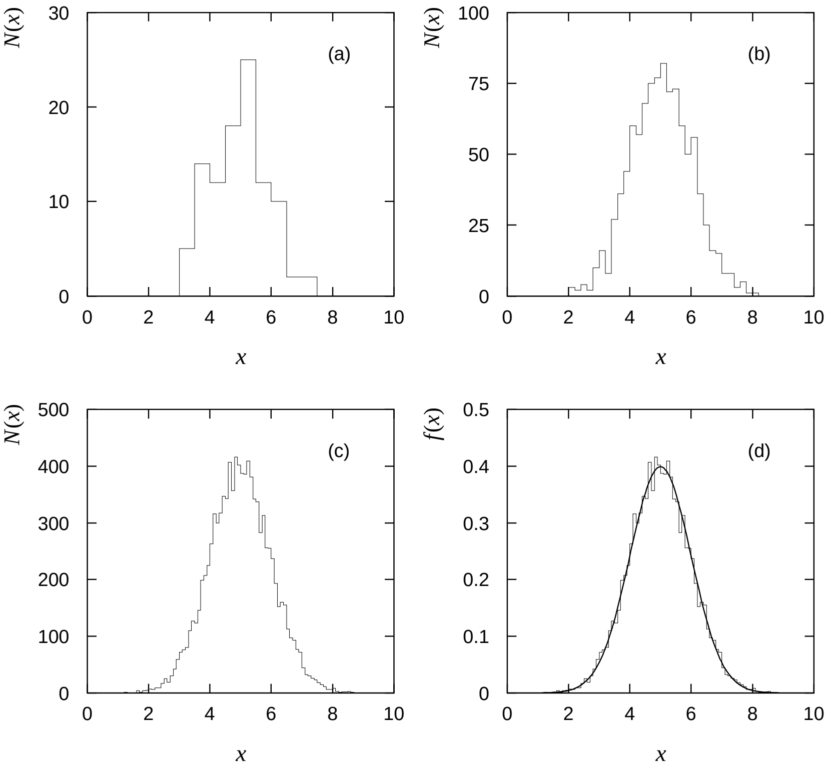

**Fig. 1.2** Histograms of various numbers of observations of a random variable $x$ based on the same p.d.f. (a) $n = 100$ observations and a bin width of $\Delta x = 0.5$. (b) $n = 1000$ observations, $\Delta x = 0.2$. (c) $n = 10000$ observations, $\Delta x = 0.1$. (d) The same histogram as in (c), but normalized to unit area. Also shown as a smooth curve is the p.d.f. according to which the observations are distributed. For (a–c), the vertical axis $N(x)$ gives the number of entries in a bin containing $x$. For (d), the vertical axis is $f(x) = N(x)/(n\Delta x)$.

One can consider cases where the variable $x$ only takes on discrete values $x_i$, for $i = 1, \ldots, N$, where $N$ can be infinite. The corresponding probabilities can be expressed as

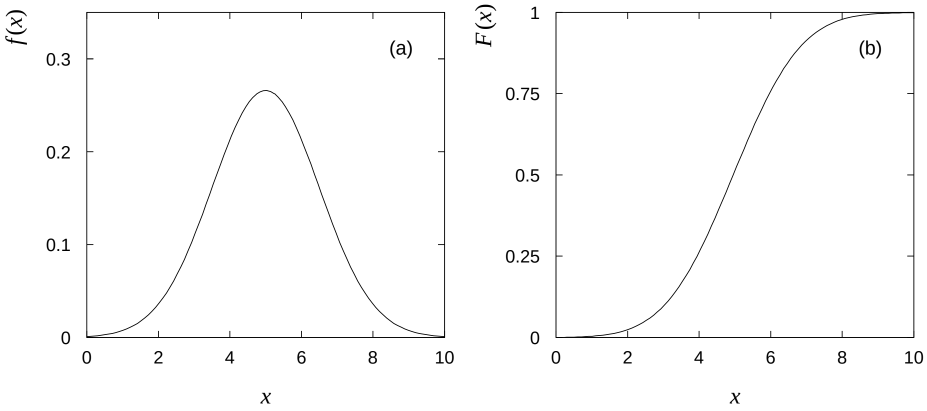

**Fig. 1.3** (a) A probability density function $f(x)$. (b) The corresponding cumulative distribution function $F(x)$.

$$\text{probability to observe value } x_i = P(x_i) = f_i, \tag{1.13}$$

where $i = 1, \ldots, N$ and the normalization condition is

$$\sum_{i=1}^{N} f_i = 1. \tag{1.14}$$

Although most of the examples in the following are done with continuous variables, the transformation to the discrete case is a straightforward correspondence between integrals and sums.

The **cumulative distribution** $F(x)$ is related to the p.d.f. $f(x)$ by

$$F(x) = \int_{-\infty}^{x} f(x')dx', \tag{1.15}$$

i.e. $F(x)$ is the probability for the random variable to take on a value less than or equal to $x$.[3] In fact, $F(x)$ is usually *defined* as the probability to obtain an outcome less than or equal to $x$, and the p.d.f. $f(x)$ is then defined as $\partial F/\partial x$. For the 'well-behaved' distributions (i.e. $F(x)$ everywhere differentiable) typically encountered in data analysis, the two approaches are equivalent. Figure 1.3 illustrates the relationship between the probability density $f(x)$ and the cumulative distribution $F(x)$.

For a discrete random variable $x_i$ with probabilities $P(x_i)$ the cumulative distribution is defined to be the probability to observe values less than or equal to the value $x$,

[3] Mathematicians call $F(x)$ the 'distribution' function, while physicists often use the word distribution to refer to the probability density function. To avoid confusion we will use the terms cumulative distribution and probability density (or p.d.f.).

$$F(x) = \sum_{x_i \leq x} P(x_i). \tag{1.16}$$

A useful concept related to the cumulative distribution is the so-called **quantile of order $\boldsymbol{\alpha}$** or **$\boldsymbol{\alpha}$-point**. The quantile $x_\alpha$ is defined as the value of the random variable $x$ such that $F(x_\alpha) = \alpha$, with $0 \leq \alpha \leq 1$. That is, the quantile is simply the inverse function of the cumulative distribution,

$$x_\alpha = F^{-1}(\alpha). \tag{1.17}$$

A commonly used special case is $x_{1/2}$, called the **median** of $x$. This is often used as a measure of the typical 'location' of the random variable, in the sense that there are equal probabilities for $x$ to be observed greater or less than $x_{1/2}$.

Another commonly used measure of location is the **mode**, which is defined as the value of the random variable at which the p.d.f. is a maximum. A p.d.f. may, of course, have local maxima. By far the most commonly used location parameter is the expectation value, which will be introduced in Section 1.5.

Consider now the case where the result of a measurement is characterized not by one but by several quantities, which may be regarded as a multidimensional random vector. If one is studying people, for example, one might measure for each person their height, weight, age, etc. Suppose a measurement is characterized by two continuous random variables $x$ and $y$. Let the event $A$ be '$x$ observed in $[x, x+dx]$ and $y$ observed anywhere', and let $B$ be '$y$ observed in $[y, y+dy]$ and $x$ observed anywhere', as indicated in Fig. 1.4.

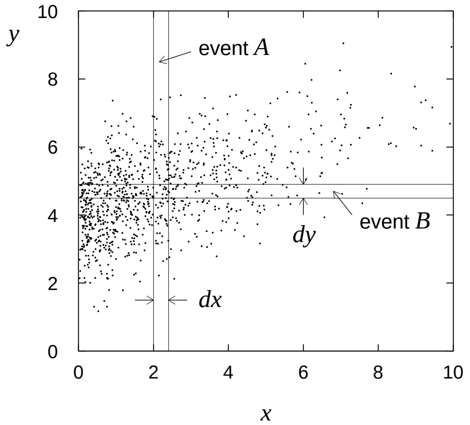

**Fig. 1.4** A scatter plot of two random variables $x$ and $y$ based on 1000 observations. The probability for a point to be observed in the square given by the intersection of the two bands (the event $A \cap B$) is given by the joint p.d.f. times the area element, $f(x,y)dxdy$.

The **joint p.d.f.** $f(x,y)$ is defined by

$$\begin{aligned} P(A \cap B) &= \text{probability of } x \text{ in } [x, x+dx] \text{ and } y \text{ in } [y, y+dy] \\ &= f(x,y)dxdy. \end{aligned} \tag{1.18}$$

The joint p.d.f. $f(x,y)$ thus corresponds to the density of points on a scatter plot of $x$ and $y$ in the limit of infinitely many points. Since $x$ and $y$ must take on some values, one has the normalization condition

$$\int\int_S f(x,y)dxdy = 1. \tag{1.19}$$

Suppose a joint p.d.f. $f(x,y)$ is known, and one would like to have the p.d.f. for $x$ regardless of the value of $y$, i.e. corresponding to event $A$ in Fig. 1.4. If one regards the 'event $A$' column as consisting of squares of area $dx\,dy$, each labeled by an index $i$, then the probability for $A$ is obtained simply by summing the probabilities corresponding to the individual squares,

$$P(A) = \sum_i f(x,y_i)dy\,dx = f_x(x)\,dx. \tag{1.20}$$

The corresponding probability density, called the **marginal p.d.f.** for $x$, is then given by the function $f_x(x)$. In the limit of infinitesimal $dy$, the sum becomes an integral, so that the marginal and joint p.d.f.s are related by

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy. \tag{1.21}$$

Similarly, one obtains the marginal p.d.f. $f_y(y)$ by integrating $f(x,y)$ over $x$,

$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx. \tag{1.22}$$

The marginal p.d.f.s $f_x(x)$ and $f_y(y)$ correspond to the normalized histograms obtained by projecting a scatter plot of $x$ and $y$ onto the respective axes. The relationship between the marginal and joint p.d.f.s is illustrated in Fig. 1.5.

From the definition of conditional probability (1.2), the probability for $y$ to be in $[y, y+dy]$ with any $x$ (event $B$) given that $x$ is in $[x, x+dx]$ with any $y$ (event $A$) is

$$P(B|A) = \frac{P(A\cap B)}{P(A)} = \frac{f(x,y)dxdy}{f_x(x)dx}. \tag{1.23}$$

The **conditional p.d.f.** for $y$ given $x$, $h(y|x)$, is thus defined as

$$h(y|x) = \frac{f(x,y)}{f_x(x)} = \frac{f(x,y)}{\int f(x,y')dy'}. \tag{1.24}$$

This is a p.d.f. of the single random variable $y$; $x$ is treated as a constant parameter. Starting from $f(x,y)$, one can simply think of holding $x$ constant, and then renormalizing the function such that its area is unity when integrated over $y$ alone.

The conditional p.d.f. $h(y|x)$ corresponds to the normalized histogram of $y$ obtained from the projection onto the $y$ axis of a thin band in $x$ (i.e. with infinitesimal width $dx$) from an $(x,y)$ scatter plot. This is illustrated in Fig. 1.6 for

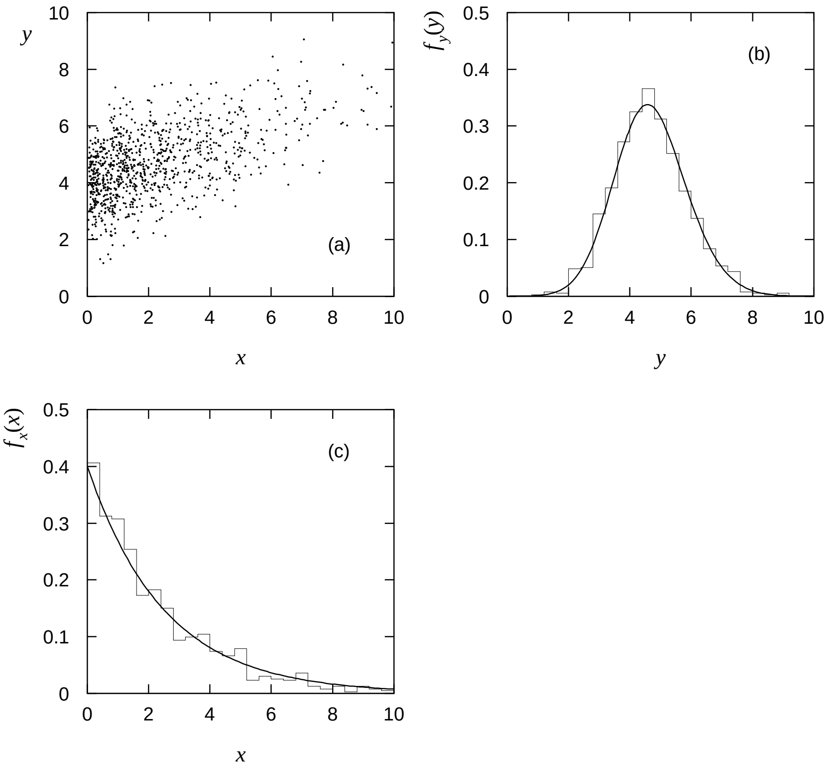

**Fig. 1.5** (a) The density of points on the scatter plot is given by the joint p.d.f. $f(x,y)$. (b) Normalized histogram from projecting the points onto the $y$ axis with the corresponding marginal p.d.f. $f_y(y)$. (c) Projection onto the $x$ axis giving $f_x(x)$.

two values of $x$, leading to two different conditional p.d.f.s, $h(y|x_1)$ and $h(y|x_2)$. Note that $h(y|x_1)$ and $h(y|x_2)$ in Fig. 1.6(b) are both normalized to unit area, as required by the definition of a probability density.

Similarly, the conditional p.d.f. for $x$ given $y$ is

$$g(x|y) = \frac{f(x,y)}{f_y(y)} = \frac{f(x,y)}{\int f(x',y)dx'}. \tag{1.25}$$

Combining equations (1.24) and (1.25) gives the relationship between $g(x|y)$ and $h(y|x)$,

$$g(x|y) = \frac{h(y|x)f_x(x)}{f_y(y)}, \tag{1.26}$$

which is Bayes' theorem for the case of continuous variables (cf. equation (1.6)).

By using $f(x,y) = h(y|x)\, f_x(x) = g(x|y)\, f_y(y)$, one can express the marginal p.d.f.s as

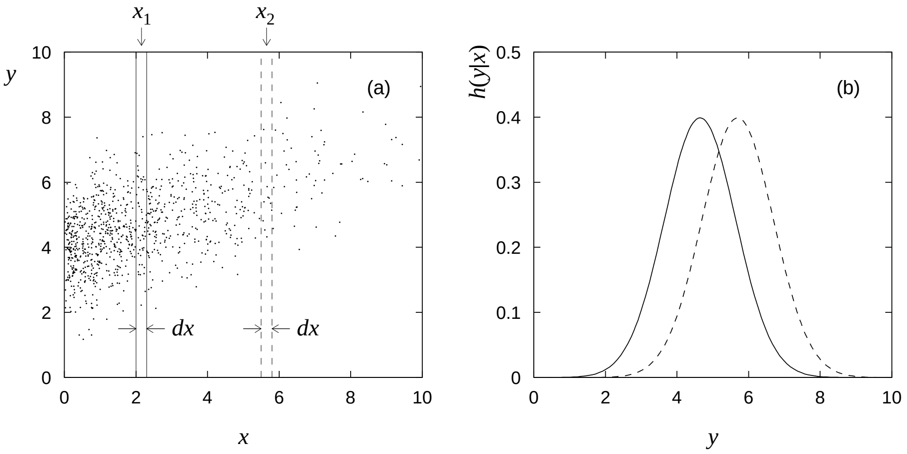

**Fig. 1.6** (a) A scatter plot of random variables $x$ and $y$ indicating two infinitesimal bands in $x$ of width $dx$ at $x_1$ (solid band) and $x_2$ (dashed band). (b) The conditional p.d.f.s $h(y|x_1)$ and $h(y|x_2)$ corresponding to the projections of the bands onto the $y$ axis.

$$f_x(x) = \int_{-\infty}^{\infty} g(x|y) f_y(y) dy, \tag{1.27}$$

$$f_y(y) = \int_{-\infty}^{\infty} h(y|x) f_x(x) dx. \tag{1.28}$$

These correspond to the law of total probability given by equation (1.7), generalized to the case of continuous random variables.

If '$x$ in $[x, x+dx]$ with any $y$' (event $A$) and '$y$ in $[y+dy]$ with any $x$' (event $B$) are independent, i.e. $P(A \cap B) = P(A)\,P(B)$, then the corresponding joint p.d.f. for $x$ and $y$ factorizes:

$$f(x, y) = f_x(x)\, f_y(y). \tag{1.29}$$

From equations (1.24) and (1.25), one sees that for independent random variables $x$ and $y$ the conditional p.d.f. $g(x|y)$ is the same for all $y$, and similarly $h(y|x)$ does not depend on $x$. In other words, having knowledge of one of the variables does not change the probabilities for the other. The variables $x$ and $y$ shown in Fig. 1.6, for example, are not independent, as can be seen from the fact that $h(y|x)$ depends on $x$.

## 1.4 Functions of random variables

Functions of random variables are themselves random variables. Suppose $a(x)$ is a continuous function of a continuous random variable $x$, where $x$ is distributed according to the p.d.f. $f(x)$. What is the p.d.f. $g(a)$ that describes the distribution of $a$? This is determined by requiring that the probability for $x$ to occur between

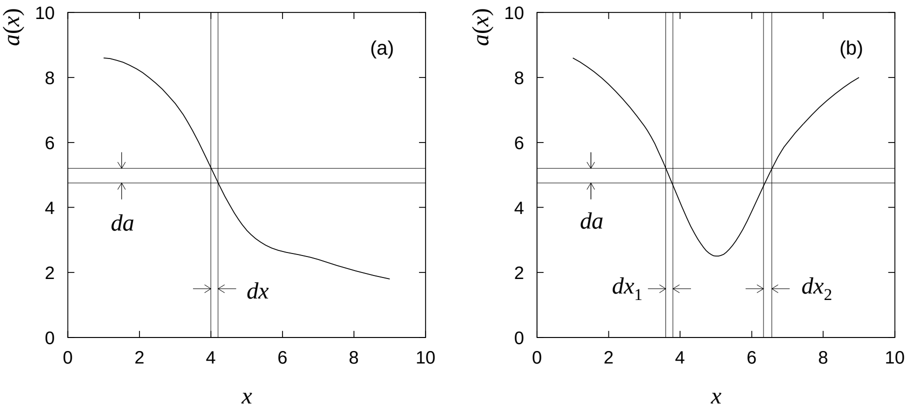

**Fig. 1.7** Transformation of variables for (a) a function $a(x)$ with a single-valued inverse $x(a)$ and (b) a function for which the interval $da$ corresponds to two intervals $dx_1$ and $dx_2$.

$x$ and $x + dx$ be equal to the probability for $a$ to be between $a$ and $a + da$. That is,

$$g(a')da' = \int_{dS} f(x)dx, \tag{1.30}$$

where the integral is carried out over the infinitesimal element $dS$ defined by the region in $x$-space between $a(x) = a'$ and $a(x) = a' + da'$, as shown in Fig. 1.7(a). If the function $a(x)$ can be inverted to obtain $x(a)$, equation (1.30) gives

$$g(a)da = \left| \int_{x(a)}^{x(a+da)} f(x')dx' \right| = \int_{x(a)}^{x(a)+\left|\frac{dx}{da}\right|da} f(x')dx', \tag{1.31}$$

or

$$g(a) = f(x(a)) \left| \frac{dx}{da} \right| . \tag{1.32}$$

The absolute value of $dx/da$ ensures that the integral is positive. If the function $a(x)$ does not have a unique inverse, one must include in $dS$ contributions from all regions in $x$-space between $a(x) = a'$ and $a(x) = a' + da'$, as shown in Fig. 1.7(b).

The p.d.f. $g(a)$ of a function $a(x_1, \ldots, x_n)$ of $n$ random variables $x_1, \ldots, x_n$ with the joint p.d.f. $f(x_1, \ldots, x_n)$ is determined by

$$g(a')da' = \int \ldots \int_{dS} f(x_1, \ldots, x_n)dx_1 \ldots dx_n, \tag{1.33}$$

where the infinitesimal volume element $dS$ is the region in $x_1, \ldots, x_n$-space between the two (hyper)surfaces defined by $a(x_1, \ldots, x_n) = a'$ and $a(x_1, \ldots, x_n) = a' + da'$.

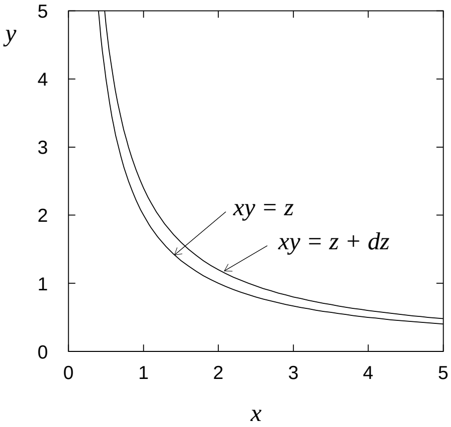

**Fig. 1.8** The region of integration $dS$ contained between the two curves $xy = z$ and $xy = z + dz$. Occurrence of $(x, y)$ values between the two curves results in occurrence of $z$ values in the corresponding interval $[z, z + dz]$.

As an example, consider two independent random variables, $x$ and $y$, distributed according to $g(x)$ and $h(y)$, and suppose we would like to find the p.d.f. of their product $z = xy$. Since $x$ and $y$ are assumed to be independent, their joint p.d.f. is given by $g(x)h(y)$. Equation (1.33) then gives for the p.d.f. of $z$, $f(z)$,

$$f(z)dz = \int\int_{dS} g(x)h(y)dxdy = \int_{-\infty}^{\infty} g(x)dx \int_{z/|x|}^{(z+dz)/|x|} h(y)dy, \tag{1.34}$$

where $dS$ is given by the region between $xy = z$ and $xy = z + dz$, as shown in Fig. 1.8. This yields

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} g(x)h(z/x)\frac{dx}{|x|} \\ &= \int_{-\infty}^{\infty} g(z/y)h(y)\frac{dy}{|y|}, \end{aligned} \tag{1.35}$$

where the second equivalent expression is obtained by reversing the order of integration. Equation (1.35) is often written $f = g \otimes h$, and the function $f$ is called the **Mellin convolution** of $g$ and $h$.

Similarly, the p.d.f. $f(z)$ of the sum $z = x + y$ is found to be

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} g(x)h(z-x)dx \\ &= \int_{-\infty}^{\infty} g(z-y)h(y)dy. \end{aligned} \tag{1.36}$$

Equation (1.36) is also often written $f = g \otimes h$, and $f$ is called the **Fourier**

**convolution** of $g$ and $h$. In the literature the names Fourier and Mellin are often dropped and one must infer from context what kind of convolution is meant.

Starting from $n$ random variables, $\mathbf{x} = (x_1, \ldots, x_n)$, the following technique can be used to determine the joint p.d.f. of $n$ linearly independent functions $a_i(\mathbf{x})$, with $i = 1, \ldots, n$. Assuming the functions $a_1, \ldots, a_n$ can be inverted to give $x_i(a_1, \ldots, a_n)$, $i = 1, \ldots, n$, the joint p.d.f. for the $a_i$ is given by

$$g(a_1, \ldots, a_n) = f(x_1, \ldots, x_n)|J|, \tag{1.37}$$

where $|J|$ is the absolute value of the Jacobian determinant for the transformation,

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial a_1} & \frac{\partial x_1}{\partial a_2} & \cdots & \frac{\partial x_1}{\partial a_n} \\ \frac{\partial x_2}{\partial a_1} & \frac{\partial x_2}{\partial a_2} & \cdots & \frac{\partial x_2}{\partial a_n} \\ \vdots & & & \vdots \\ & & \cdots & \frac{\partial x_n}{\partial a_n} \end{vmatrix} . \tag{1.38}$$

To determine the marginal p.d.f. for one of the functions (say $g_1(a_1)$) the joint p.d.f. $g(a_1, \ldots, a_n)$ must be integrated over the remaining $a_i$.

In many cases the techniques given above are too difficult to solve analytically. For example, if one is interested in a single function of $n$ random variables, where $n$ is some large and itself possibly variable number, it is rarely practical to come up with $n-1$ additional functions and then integrate the transformed joint p.d.f. over the unwanted ones. In such cases a numerical solution can usually be found using the Monte Carlo techniques discussed in Chapter 3. If only the mean and variance of a function are needed, the so-called 'error propagation' procedures described in Section 1.6 can be applied.

For certain cases the p.d.f. of a function of random variables can be found using integral transform techniques, specifically, Fourier transforms of the p.d.f.s for sums of random variables and Mellin transforms for products. The basic idea is to take the Mellin or Fourier transform of equation (1.35) or (1.36), respectively. The equation $f = g \otimes h$ is then converted into the product of the transformed density functions, $\tilde{f} = \tilde{g} \cdot \tilde{h}$. The p.d.f. $f$ is obtained by finding the inverse transform of $\tilde{f}$. A complete discussion of these methods is beyond the scope of this book; see e.g. [Spr79]. Some examples of sums of random variables using Fourier transforms (characteristic functions) are given in Chapter 10.

## 1.5 Expectation values

The **expectation value** $E[x]$ of a random variable $x$ distributed according to the p.d.f. $f(x)$ is defined as

$$E[x] = \int_{-\infty}^{\infty} x f(x) dx = \mu. \tag{1.39}$$

The expectation value of $x$ (also called the **population mean** or simply the mean of $x$) is often denoted by $\mu$. Note that $E[x]$ is not a function of $x$, but depends rather on the form of the p.d.f. $f(x)$. If the p.d.f. $f(x)$ is concentrated mostly in one region, then $E[x]$ represents a measure of where values of $x$ are likely to be observed. It can be, however, that $f(x)$ consists of two widely separated peaks, such that $E[x]$ is in the middle where $x$ is seldom (or never) observed.

For a function $a(x)$, the expectation value is

$$E[a] = \int_{-\infty}^{\infty} a g(a) da = \int_{-\infty}^{\infty} a(x) f(x) dx, \tag{1.40}$$

where $g(a)$ is the p.d.f. of $a$ and $f(x)$ is the p.d.f. of $x$. The second integral is equivalent; this can be seen by multiplying both sides of equation (1.30) by $a$ and integrating over the entire space.

Some more expectation values of interest are:

$$E[x^n] = \int_{-\infty}^{\infty} x^n f(x) dx = \mu_n', \tag{1.41}$$

called the $n$th algebraic moment of $x$, for which $\mu = \mu_1'$ is a special case, and

$$E[(x - E[x])^n] = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx = \mu_n, \tag{1.42}$$

called the $n$th central moment of $x$. In particular, the second central moment,

$$E[(x - E[x])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \sigma^2 = V[x], \tag{1.43}$$

is called the **population variance** (or simply the variance) of $x$, written $\sigma^2$ or $V[x]$. Note that $E[(x - E[x])^2] = E[x^2] - \mu^2$. The variance is a measure of how widely $x$ is spread about its mean value. The square root of the variance $\sigma$ is called the **standard deviation** of $x$, which is often useful because it has the same units as $x$.

For the case of a function $a$ of more than one random variable $\mathbf{x} = (x_1, \ldots, x_n)$, the expectation value is

$$\begin{aligned} E[a(\mathbf{x})] &= \int_{-\infty}^{\infty} a g(a) da \\ &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} a(\mathbf{x}) f(\mathbf{x}) dx_1 \ldots dx_n = \mu_a, \end{aligned} \tag{1.44}$$

where $g(a)$ is the p.d.f. for $a$ and $f(\mathbf{x})$ is the joint p.d.f. for the $x_i$. In the following, the notation $\mu_a = E[a]$ will often be used. As in the single-variable case, the two integrals in (1.44) are equivalent, as can be seen by multiplying both sides of equation (1.33) by $a$ and integrating over the entire space. The variance of $a$ is

$$\begin{aligned} V[a] &= E[(a-\mu_a)^2] \\ &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (a(\mathbf{x})-\mu_a)^2 f(\mathbf{x}) dx_1 \ldots dx_n = \sigma_a^2, \end{aligned} \qquad (1.45)$$

and is denoted by $\sigma_a^2$ or $V[a]$. The **covariance** of two random variables $x$ and $y$ is defined as

$$\begin{aligned} V_{xy} &= E[(x-\mu_x)(y-\mu_y)] = E[xy] - \mu_x \mu_y \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, y\, f(x,y)\, dx\, dy \;-\; \mu_x \mu_y, \end{aligned} \qquad (1.46)$$

where $\mu_x = E[x]$ and $\mu_y = E[y]$. The covariance matrix $V_{xy}$, also called the error matrix, is sometimes denoted by $\text{cov}[x,y]$. More generally, for two functions $a$ and $b$ of $n$ random variables $\mathbf{x} = (x_1, \ldots, x_n)$, the covariance $\text{cov}[a,b]$ is given by

$$\begin{aligned} \text{cov}[a,b] &= E[(a-\mu_a)(b-\mu_b)] \\ &= E[ab] - \mu_a \mu_b \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a\, b\, g(a,b)\, da\, db \;-\; \mu_a \mu_b \\ &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} a(\mathbf{x})\, b(\mathbf{x})\, f(\mathbf{x}) dx_1 \ldots dx_n \;-\; \mu_a \mu_b, \end{aligned} \qquad (1.47)$$

where $g(a,b)$ is the joint p.d.f. for $a$ and $b$ and $f(\mathbf{x})$ is the joint p.d.f. for the $x_i$. As in equation (1.44), the two integral expressions for $V_{ab}$ are equivalent. Note that by construction the covariance matrix $V_{ab}$ is symmetric in $a$ and $b$ and that the diagonal elements $V_{aa} = \sigma_a^2$ (i.e. the variances) are positive.

In order to give a dimensionless measure of the level of correlation between two random variables $x$ and $y$, one often uses the **correlation coefficient**, defined by

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}. \qquad (1.48)$$

One can show (see e.g. [Fro79, Bra92]) that the correlation coefficient lies in the range $-1 \le \rho_{xy} \le 1$.

One can roughly understand the covariance of two random variables $x$ and $y$ in the following way. $V_{xy}$ is the expectation value of $(x-\mu_x)(y-\mu_y)$, the product of the deviations of $x$ and $y$ from their means, $\mu_x$ and $\mu_y$. Suppose that

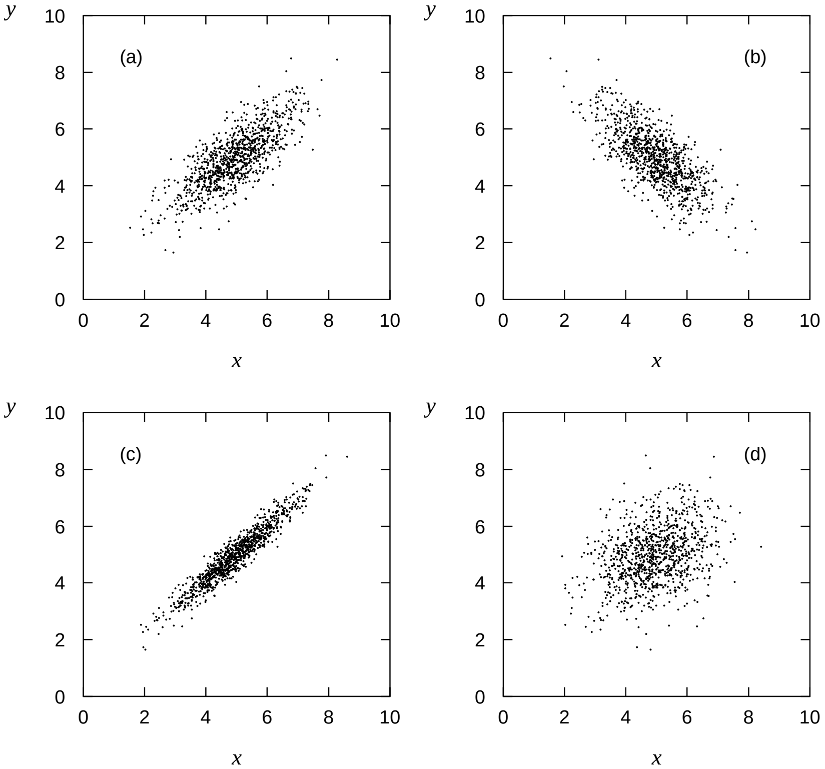

**Fig. 1.9** Scatter plots of random variables $x$ and $y$ with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of $x$ and $y$ are $\sigma_x = \sigma_y = 1$.

having $x$ greater than $\mu_x$ enhances the probability to find $y$ greater than $\mu_y$, and $x$ less than $\mu_x$ gives an enhanced probability to have $y$ less than $\mu_y$. Then $V_{xy}$ is greater than zero, and the variables are said to be positively correlated. Such a situation is illustrated in Figs 1.9 (a), (c) and (d), for which the correlation coefficients $\rho_{xy}$ are 0.75, 0.95 and 0.25, respectively. Similarly, $V_{xy} < 0$ is called a negative correlation: having $x > \mu_x$ increases the probability to observe $y < \mu_y$. An example is shown in Fig. 1.9(b), for which $\rho_{xy} = -0.75$.

From equations (1.29) and (1.44), it follows that for independent random variables $x$ and $y$,

$$E[xy] = E[x]E[y] = \mu_x \mu_y, \tag{1.49}$$

(and hence by equation (1.46), $V_{xy} = 0$) although the converse is not necessarily true. Figure 1.10, for example, shows a two-dimensional scatter plot of a p.d.f. for which $V_{xy} = 0$, but where $x$ and $y$ are not independent. That is, $f(x, y)$ does not factorize according to equation (1.29), and hence knowledge of one of the

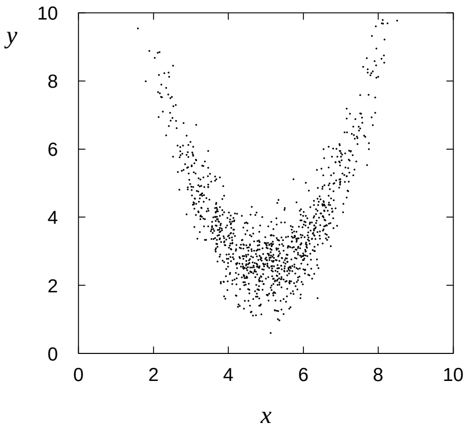

**Fig. 1.10** Scatter plot of random variables $x$ and $y$ which are not independent (i.e. $f(x,y) \neq f_x(x)f_y(y)$) but for which $V_{xy} = 0$ because of the particular symmetry of the distribution.

variables affects the conditional p.d.f. of the other. The covariance $V_{xy}$ vanishes, however, because $f(x,y)$ is symmetric in $x$ about the mean $\mu_x$.

## 1.6 Error propagation

Suppose one has a set of $n$ random variables $\mathbf{x} = (x_1, \ldots, x_n)$ distributed according to some joint p.d.f. $f(\mathbf{x})$. Suppose that the p.d.f. is not completely known, but the mean values of the $x_i$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, and the covariance matrix, $V_{ij}$, are known or have at least been estimated. (Methods for doing this are described in Chapter 5.)

Now consider a function of the $n$ variables $y(\mathbf{x})$. To determine the p.d.f. for $y$, one must in principle follow a procedure such as those described in Section 1.4 (e.g. equations (1.33) or (1.37)). We have assumed, however, that $f(\mathbf{x})$ is not completely known, only the means $\boldsymbol{\mu}$ and the covariance matrix $V_{ij}$, so this is not possible. One can, however, approximate the expectation value of $y$ and the variance $V[y]$ by first expanding the function $y(\mathbf{x})$ to first order about the mean values of the $x_i$,

$$y(\mathbf{x}) \approx y(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} (x_i - \mu_i). \tag{1.50}$$

The expectation value of $y$ is to first order

$$E[y(\mathbf{x})] \approx y(\boldsymbol{\mu}), \tag{1.51}$$

since $E[x_i - \mu_i] = 0$. The expectation value of $y^2$ is

$$\begin{aligned} E[y^2(\mathbf{x})] &\approx y^2(\boldsymbol{\mu}) + 2y(\boldsymbol{\mu}) \cdot \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} E[x_i - \mu_i] \\ &+ E\left[ \left( \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} (x_i - \mu_i) \right) \left( \sum_{j=1}^{n} \left[ \frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} (x_j - \mu_j) \right) \right] \\ &= y^2(\boldsymbol{\mu}) + \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} V_{ij}, \end{aligned} \tag{1.52}$$

so that the variance $\sigma_y^2 = E[y^2] - (E[y])^2$ is given by

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} V_{ij}. \tag{1.53}$$

Similarly, one obtains for a set of $m$ functions $y_1(\mathbf{x}), \ldots, y_m(\mathbf{x})$ the covariance matrix

$$U_{kl} = \mathrm{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[ \frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} V_{ij}. \tag{1.54}$$

This can be expressed in matrix notation as

$$U = A\, V\, A^T, \tag{1.55}$$

where the matrix of derivatives $A$ is

$$A_{ij} = \left[ \frac{\partial y_i}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} \tag{1.56}$$

and $A^T$ is the transpose of $A$. Equations (1.53)–(1.56) form the basis of **error propagation** (i.e. the variances, which are used as measures of statistical uncertainties, are propagated from the $x_i$ to the functions $y_1$, $y_2$, etc.). (The term 'error' will often be used to refer to the uncertainty of a measurement, which in most cases is given by the standard deviation of the corresponding random variable.)

For the case where the $x_i$ are not correlated, i.e. $V_{ii} = \sigma_i^2$ and $V_{ij} = 0$ for $i \neq j$, equations (1.53) and (1.54) become

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]^2_{\mathbf{x}=\boldsymbol{\mu}} \sigma_i^2 \tag{1.57}$$

and

$$U_{kl} \approx \sum_{i=1}^{n} \left[ \frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} \sigma_i^2. \tag{1.58}$$

Equation (1.53) leads to the following special cases. If $y = x_1 + x_2$, the variance of $y$ is then

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}. \tag{1.59}$$

For the product $y = x_1 x_2$ one obtains

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{V_{12}}{x_1 x_2}. \tag{1.60}$$

If the variables $x_1$ and $x_2$ are not correlated ($V_{12} = 0$), the relations above state that errors (i.e. standard deviations) add quadratically for the sum $y = x_1 + x_2$, and that the *relative* errors add quadratically for the product $y = x_1 x_2$.

In deriving the error propagation formulas we have assumed that the means and covariances of the original set of variables $x_1, \ldots, x_n$ are known (or at least estimated) and that the desired functions of these variables can be approximated by the first-order Taylor expansion around the means $\mu_1, \ldots, \mu_n$. The latter assumption is of course only exact for a linear function. The approximation breaks down if the function $y(\mathbf{x})$ (or functions $\mathbf{y}(\mathbf{x})$) are significantly nonlinear in a region around the means $\boldsymbol{\mu}$ of a size comparable to the standard deviations of the $x_i$, $\sigma_1, \ldots, \sigma_n$. Care must be taken, for example, with functions like $y(x) = 1/x$ when $E[x] = \mu$ is comparable to or smaller than the standard deviation of $x$. Such situations can be better treated with the Monte Carlo techniques described in Chapter 3, or using confidence intervals as described in Section 9.2.

## 1.7 Orthogonal transformation of random variables

Suppose one has a set of $n$ random variables $x_1, \ldots, x_n$ and their covariance matrix $V_{ij} = \mathrm{cov}[x_i, x_j]$, for which the off-diagonal elements are not necessarily zero. Often it can be useful to define $n$ new variables $y_1, \ldots, y_n$ that are not correlated, i.e. for which the new covariance matrix $U_{ij} = \mathrm{cov}[y_i, y_j]$ is diagonal. We will show that this is always possible with a linear transformation,

$$y_i = \sum_{j=1}^{n} A_{ij} x_j. \tag{1.61}$$

Assuming such a transformation, the covariance matrix for the new variables is

$$
\begin{aligned}
U_{ij} = \mathrm{cov}[y_i, y_j] &= \mathrm{cov}\left[\sum_{k=1}^{n} A_{ik} x_k, \sum_{l=1}^{n} A_{jl} x_l\right] \\
&= \sum_{k,l=1}^{n} A_{ik} A_{jl} \,\mathrm{cov}[x_k, x_l] \\
&= \sum_{k,l=1}^{n} A_{ik} V_{kl} A_{lj}^T . \qquad (1.62)
\end{aligned}
$$

This is simply a special case of the error propagation formula (1.54); here it is exact, since the function (1.61) is linear.

The problem thus consists of finding a matrix $A$ such that $U = AVA^T$ is diagonal. This is simply the diagonalization of a real, symmetric matrix, a well-known problem of linear algebra (cf. [Arf95]). The solution can be found by first determining the eigenvectors $\mathbf{r}^i$, $i = 1, \ldots, n$, of the covariance matrix $V$. That is, one must solve the equation

$$V\mathbf{r}^i = \lambda_i \mathbf{r}^i , \qquad (1.63)$$

where in the matrix equations the vector $\mathbf{r}$ should be understood as a column vector. The eigenvectors $\mathbf{r}^i$ are only determined up to a multiplicative factor, which can be chosen such that they all have unit length. Furthermore, one can easily show that since the covariance matrix is symmetric, the eigenvectors are orthogonal, i.e.

$$\mathbf{r}^i \cdot \mathbf{r}^j = \sum_{k=1}^{n} r_k^i r_k^j = \delta_{ij} . \qquad (1.64)$$

If two or more of the eigenvalues $\lambda_i, \lambda_j, \ldots$ are equal, then the directions of the corresponding eigenvectors $\mathbf{r}^i, \mathbf{r}^j, \ldots$ are not uniquely determined, but can nevertheless be chosen such that the eigenvectors are orthogonal.

The $n$ rows of the transformation matrix $A$ are then given by the $n$ eigenvectors $\mathbf{r}^i$ (in any order), i.e. $A_{ij} = r_j^i$, and the transpose matrix thus has the eigenvectors as its columns, $A_{ij}^T = r_i^j$. That this matrix has the desired property can be shown explicitly by substituting it into equation (1.62),

$$
\begin{aligned}
U_{ij} = \sum_{k,l=1}^{n} A_{ik} V_{kl} A_{lj}^T &= \sum_{k,l=1}^{n} r_k^i V_{kl} r_l^j \\
&= \sum_{k=1}^{n} r_k^i \lambda_j r_k^j \\
&= \lambda_j \mathbf{r}^i \cdot \mathbf{r}^j \\
&= \lambda_j \delta_{ij} . \qquad (1.65)
\end{aligned}
$$

Thus the variances of the transformed variables $y_1, \ldots, y_n$ are given by the eigenvalues of the original covariance matrix $V$, and all off-diagonal elements of $U$ are zero. Since the eigenvectors are orthonormal (equation (1.64)), one has the property

$$\sum_{j=1}^{n} A_{ij} A^T_{jk} = \sum_{j=1}^{n} r^i_j r^k_j = \mathbf{r}^i \cdot \mathbf{r}^k = \delta_{ik}, \tag{1.66}$$

or as a matrix equation $AA^T = 1$, and hence $A^T = A^{-1}$. Such a transformation is said to be orthogonal, i.e. it corresponds to a rotation of the vector $\mathbf{x}$ into $\mathbf{y}$ such that the norm remains constant, since $|\mathbf{y}|^2 = \mathbf{y}^T\mathbf{y} = \mathbf{x}^T A^T A \mathbf{x} = |\mathbf{x}|^2$.

In order to find the eigenvectors of $V$, the standard techniques of linear algebra can used (see e.g. [Arf95]). For more than three variables, the problem becomes impractical to solve analytically, and numerical techniques such as the **singular value decomposition** are necessary (see e.g. [Bra92, Pre92]).

In two dimensions, for example, the covariance matrix for the variables $\mathbf{x} = (x_1, x_2)$ can be expressed as

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{1.67}$$

The eigenvalue equation $(V - I\lambda)\mathbf{r} = 0$ (where $I$ is the $2 \times 2$ unit matrix) is solved by requiring that the determinant of the matrix of coefficients be equal to zero,

$$\det(V - I\lambda) = 0. \tag{1.68}$$

The two eigenvalues $\lambda_\pm$ are found to be

$$\lambda_\pm = \tfrac{1}{2}\left[\sigma_1^2 + \sigma_2^2 \pm \sqrt{(\sigma_1^2 + \sigma_2^2)^2 - 4(1-\rho^2)\sigma_1^2\sigma_2^2}\right]. \tag{1.69}$$

The two orthonormal eigenvectors $\mathbf{r}_\pm$ can be parametrized by an angle $\theta$,

$$\mathbf{r}_+ = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \qquad \mathbf{r}_- = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}. \tag{1.70}$$

Substituting the eigenvalues (1.69) back into the eigenvalue equation determines the angle $\theta$,

$$\theta = \tfrac{1}{2}\tan^{-1}\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right). \tag{1.71}$$

The rows of the desired transformation matrix are thus given by the two eigenvectors,

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \ . \tag{1.72}$$

This corresponds to a rotation of the vector $(x_1, x_2)$ by an angle $\theta$. An example is shown in Fig. 1.11 where the original two variables have $\sigma_1 = 1.5$, $\sigma_2 = 1.0$, and a correlation coefficient of $\rho = 0.7$.

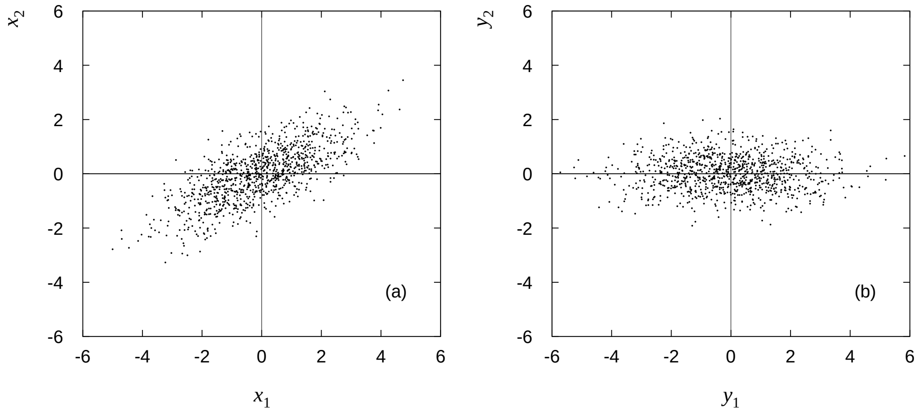

**Fig. 1.11** Scatter plot of (a) two correlated random variables $(x_1, x_2)$ and (b) the transformed variables $(y_1, y_2)$ for which the covariance matrix is diagonal.

Although uncorrelated variables are often easier to deal with, the transformed variables may not have as direct an interpretation as the original ones. Examples where this procedure could be used will arise in Chapters 6 through 8 on parameter estimation, where the estimators for a set of parameters will often be correlated.

# 2

# Examples of probability functions

In this chapter a number of commonly used probability distributions and density functions are presented. Properties such as mean and variance are given, mostly without proof; the moments can be found by using characteristic functions introduced in Chapter 10. Additional p.d.f.s can be found in [Fro79] Chapter 4, [Ead71] Chapter 4, [Bra92] Chapter 5.

## 2.1 Binomial and multinomial distributions

Consider a series of $N$ independent trials or observations, each having two possible outcomes, here called 'success' and 'failure', where the probability for success is some constant value, $p$. The set of trials can be regarded as a single measurement and is characterized by a discrete random variable $n$, defined to be the total number of successes. That is, the sample space is defined to be the set of possible values of $n$ successes given $N$ observations. If one were to repeat the entire experiment many times with $N$ trials each time, the resulting values of $n$ would occur with relative frequencies given by the so-called **binomial distribution**.

The form of the binomial distribution can be derived in the following way. We have assumed that the probability of success in a single observation is $p$ and the probability of failure is $1 - p$. Since the individual trials are assumed to be independent, the probability for a series of successes and failures in a particular order is equal to the product of the individual probabilities. For example, the probability in five trials to have success, success, failure, success, failure in that order is $p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3(1-p)^2$. In general the probability for a particular sequence of $n$ successes and $N - n$ failures is $p^n(1-p)^{N-n}$. We are not interested in the order, however, only in the final number of successes $n$. The number of sequences having $n$ successes in $N$ events is

$$\frac{N!}{n!(N-n)!}, \tag{2.1}$$

so the total probability to have $n$ successes in $N$ events is

$$f(n; N, p) = \frac{N!}{n!(N-n)!} \, p^n \, (1-p)^{N-n}, \tag{2.2}$$