1

Probability

1.1 The Laws of Probability

Figure 1.1 shows a standard six-sided die, each side distinguished by a unique number of dots from 1 to 6. Consider these three questions:

1. The die is thrown and lands with one side up. Before we look at the die and know the answer, we ask "What is the probability that the side with three dots is up?"
2. The die is weighed many times, each time by a different scale, and the weights are recorded. "How much does the die weigh, and how reliable is the measurement of the weight?"
3. The die is thrown many times, and the up side is recorded each time. Let H be the hypothesis that all faces are equally likely to be on the up side. "What is the probability that H is true?"

Probability is a deductive discipline relying on pre-existing information. Statistics deals with observations of the real world and inferences made from them. The first question can be answered before actually throwing the die if enough is known about the die. It lies in the province of probability. The answers to the second and third questions depend on the measurements. They lie in the province of statistics. The subject of this book is data analysis, which lies firmly in the province of statistics; but the language and mathematical tools of statistics rely heavily on probability. We begin, therefore, with a discussion of probability.

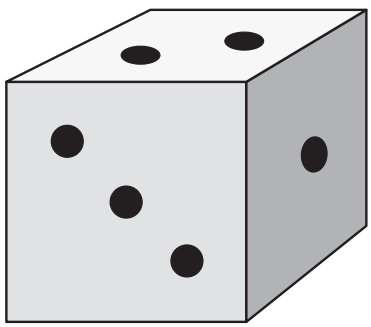

Figure 1.1: A typical six-sided die.

The classical definition of probability is based on frequency. Suppose that an event A can happen in k ways out of a total of n possible, equally likely, ways. The probability that A will occur is

$$P(A) = \frac{k}{n}. \tag{1.1}$$

This definition presupposes that enough is known to permit the calculation of k and n in advance, so it is sometimes called the *a priori* probability. Although the definition may seem reasonable, it has several limitations. First, both k and n must be finite. The answer to the question "If one picks out a star in the universe at random, what is the probability that it is a neutron star?" cannot be calculated from the classical definition if the universe is infinite in size. If k and n are not finite, it is better to define probability as the limit

$$P(A) = \lim_{n\to\infty} \frac{k}{n}. \tag{1.2}$$

Next, k and n must both be countable. Consider an archer shooting an arrow at a target. The answer to the question "What is the probability that an arrow will hit within 1 inch of the center of a target?" cannot be calculated from equations 1.1 and 1.2, because the possible places the arrow might hit form a continuum and are not countable. This problem can be fixed by replacing k and n with integrals over regions. The probability that A occurs is the fraction of the regions in which event A occurs, producing yet a third definition of probability. These three definitions are, of course, closely related. We will need to use all three.

Another problem with the classical definition of probability is that the words "equally likely" actually mean "equally probable," so the definition is circular! One must add an independent set of rules for calculating k and n, or at least for assessing whether probabilities for two events are equal. If the rules are chosen cleverly, the calculations will yield results in accord with our intuition about probability; but there is no guarantee the rules have been chosen correctly and certainly no guarantee that the calculations will apply to our universe. In practice one breaks the circularity by invoking external arguments or information. One might, for example, claim that a six-sided die is symmetric, so each side has the same probability of landing face up.

Finally, we shall see that Bayesian statistics allows calculation of probabilities for unique events. Practitioners of Bayesian statistics can and do calculate the probability that, for example, a specific person will be elected president of the United States in a specific election year. Frequency is meaningless for unique events, so Bayesian statistics requires a reassessment of the meaning of probability. We will delay a discussion of nonfrequentist interpretations of probability until Chapter 7.

One typically calculates probabilities for a complicated problem by breaking it into simpler parts whose probabilities are more easily calculated and then using the laws of probability to combine the probabilities for the simple parts into probabilities for the more complicated problem. The laws of probability can be derived with the help of the Venn diagrams shown in Figure 1.2. Let S be the set of all possible outcomes of an experiment and let A be a subset of S, denoted by

$$A \subset S \tag{1.3}$$

(see the top Venn diagram in Figure 1.2). If the outcomes can be counted and the number of outcomes is finite, then

$$P(A) = \frac{n_A}{n_S}, \tag{1.4}$$

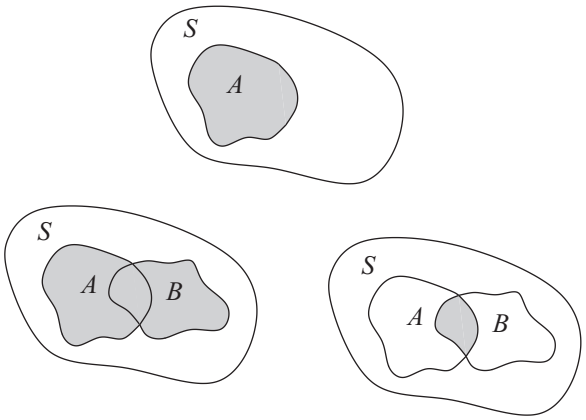

Figure 1.2: The upper figure shows the subset A of the set S. The two lower figures include a second subset B. The shaded region in the left-hand figure is $A \cup B$, and the shaded region in the right-hand figure is $A \cap B$.

where n_S is the total number of outcomes in S, and n_A is the number of outcomes in the subset A. If outcomes in A never occur, so that A is an empty set, then $P(A) = 0/n_S = 0$. If A includes all of S, so that all outcomes are in S, then $n_A = n_S$, and $P(A) = n_S/n_S = 1$. The probability must, then, lie in the range

$$0 \leq P(A) \leq 1. \tag{1.5}$$

The intersection and union of two sets are denoted by the symbols $\cap$ and $\cup$:

$A \cup B = B \cup A =$ the union of sets A and B with the overlap counted just once,
$A \cap B = B \cap A =$ the intersection of A and B.

The meanings of these operations are shown by the two lower Venn diagrams in Figure 1.2. Suppose we have two subsets of S, A and B. The probability that an outcome lies in A or B or both is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.6}$$

The first two terms on the right-hand side are simply the probabilities that the outcome lies in either A or B. We cannot just add the two probabilities together to get $P(A \cup B)$, because the two sets might have outcomes in common—the two sets might overlap. If they do, the overlap region is counted twice by the first two terms. The third term is the overlap of the two sets. Subtracting it removes one of the double-counted overlap regions. If the two sets do not intersect, so that $P(A \cap B) = 0$, then equation 1.6 reduces to

$$P(A \cup B) = P(A) + P(B). \tag{1.7}$$

The complement of A, denoted by $\bar{A}$, is the subset of S consisting of all members of S not in A (see Figure 1.3). Together A and $\bar{A}$ make up all of S, so

$$P(A \cup \bar{A}) = P(S). \tag{1.8}$$

They do not overlap, so

$$P(A \cap \bar{A}) = 0. \tag{1.9}$$

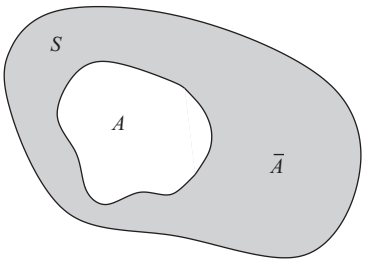

Figure 1.3: The subset $\bar{A}$ of the set S consists of all members of S not in A.

Therefore,

$$1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A}), \tag{1.10}$$

from which we find

$$P(\bar{A}) = 1 - P(A). \tag{1.11}$$

The conditional probability, denoted by $P(A|B)$, is the probability that A occurs if B occurs, often stated as the probability of A "given" B to avoid implications of causality. It is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.12}$$

Example: Suppose we were to throw two dice and the sum of their faces is 7. What is the probability that one of the faces is a 5? In other words, what is the conditional probability $P(5|\text{sum} = 7)$?

There are 36 possible ways the faces of the two dice can land up. The faces can add to 7 in the following 6 ways:

1-6	4-3
2-5 X	5-2 X
3-4	6-1

Therefore $P(\text{sum} = 7) = 6/36 = 1/6$. Two of the ways of adding to 7, the ones marked with an X, contain a 5. So, the probability that the sum is 7 and also that one die is a 5 is $P(5 \cap \text{sum} = 7) = 2/36 = 1/18$. The conditional probability is, therefore,

$$P(5|\text{sum} = 7) = \frac{P(5 \cap \text{sum} = 7)}{P(\text{sum} = 7)} = \frac{1/18}{1/6} = 1/3$$

We could, of course, have calculated the conditional probability directly by noting that 1/3 of the entries in the list of combinations are marked with an X.

Events A and B are said to be independent of each other if the occurrence of one event does not affect the probability that the other occurs. This is expressed by

$$P(A|B) = P(A). \tag{1.13}$$

Plugging equation 1.13 into equation 1.12, we derive another way of expressing independence:

$$P(A \cap B) = P(A)P(B). \tag{1.14}$$

So, if events A and B are independent of each other, the probability that both occur is given by the products of the probabilities that each occurs.

Because of the symmetry between A and B, there are two ways to write $P(A \cap B)$:

$$P(A \cap B) = P(A|B)\,P(B) \tag{1.15}$$

and

$$P(A \cap B) = P(B|A)\,P(A). \tag{1.16}$$

Together, these two equations mean

$$P(A|B)\,P(B) = P(B|A)\,P(A) \tag{1.17}$$

and, therefore,

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}. \tag{1.18}$$

This important result is the basis of Bayesian statistics.

1.2 Probability Distributions

1.2.1 Discrete and Continuous Probability Distributions

Up to now we have assumed that an experiment has two possible outcomes, A or B. Now let there be many possible outcomes. We denote the outcomes by A_j, where j is an integer index, and the probability of outcome A_j by $P(A_j)$. To be a valid probability distribution, $P(A_j)$ must be single valued and satisfy

$$P(A_j) \geq 0 \tag{1.19}$$

$$\sum_j P(A_j) = 1. \tag{1.20}$$

These mild constraints leave wide latitude for possible discrete probability distributions. The other laws of probability generalize in obvious ways. One is worth mentioning explicitly since it occurs so often: If n events A_j are independent of one another, the probability that all will occur is equal to the product of the individual probabilities of occurrence:

$$P(A_1 \cap A_2 \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n). \tag{1.21}$$

It is also worth working out one nontrivial example for overlapping probabilities: The probability that any of three possible outcomes occur is

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P((A_1 \cup A_2) \cup A_3) \\ &= P(A_1 \cup A_2) + P(A_3) - P((A_1 \cup A_2) \cap A_3) \\ &= P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) \\ &\quad - [P(A_1 \cap A_3) + P(A_2 \cap A_3) - P(A_1 \cap A_2 \cap A_3)] \\ &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &\quad + P(A_1 \cap A_2 \cap A_3). \end{aligned} \tag{1.22}$$

The following example shows how these rules can be used to calculate a priori probabilities for a symmetric six-sided die.

Example: Calculate from first principles the probability that any one face of a six-sided die will land face up.

From the physical properties of dice, one or another face must land face up. Therefore probability that the die lands with a 1 or 2 or 3 or 4 or 5 or 6 on the top face must equal 1. We express this by the equation

$$P(1 \cup 2 \cup 3 \cup 4 \cup 5 \cup 6) = 1.$$

Again from physical properties of dice, only one of the faces can land face up. This requires that all the intersections are equal to zero:

$$P(1 \cap 2) = P(1 \cap 3) = \cdots = P(5 \cap 6) = 0.$$

Extending equation 1.22 to six possible outcomes, we have

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1.$$

If the die is nearly symmetric, the probabilities must all be equal:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6).$$

We have, then

$$6P(1) = 1$$

$$P(1) = \frac{1}{6},$$

and the other faces have the same probability of landing face up.

The outcomes of an experiment can also form a continuum. Let each outcome be any real number x. A function $\mathrm{f}(x)$ can be a probability distribution function if it is single-valued and if

$$\mathrm{f}(x) \geq 0 \tag{1.23}$$

$$\int_{-\infty}^{\infty} \mathrm{f}(x)\, dx = 1 \tag{1.24}$$

$$P(a \leq x \leq b) = \int_a^b \mathrm{f}(x)\, dx, \tag{1.25}$$

where $P(a \leq x \leq b)$ is the probability that x lies in the range $a \leq x \leq b$. Equation 1.23 ensures that all probabilities are positive, and equation 1.24 ensures that the probability of all possible outcomes is equal to 1. A function that satisfies equation 1.24 is said to be *normalized*. Together these two equations replace equations 1.19 and 1.20 for discrete probabilities. The third requirement defines the relation between $\mathrm{f}(x)$ and probability and replaces equation 1.4.

The rectangle function is a simple example of a valid continuous probability distribution function:

$$\mathrm{f}(x) = \begin{cases} 0, & x < 0 \\ 1/a, & 0 \le x \le a \\ 0, & x > a \end{cases} . \tag{1.26}$$

Note that $\mathrm{f}(x) > 1$ if $a < 1$. Since probabilities must be less than 1, $\mathrm{f}(x)$ is clearly not itself a probability. The probability is the integral of $\mathrm{f}(x)$ between two limits (equation 1.25), so one must always discuss the probability that x lies in a given range. Because of this, $\mathrm{f}(x)$ is sometimes called the *probability density distribution function*, not the probability distribution function. For example, one should properly speak of the rectangular probability density distribution function, not the rectangular probability distribution function. In practice the distinction is rarely emphasized, and the word "density" remains unverbalized.

Continuous probability distribution functions need not be continuous everywhere, and they need not even be finite everywhere! The Dirac delta function $\delta(x)$ is a legitimate probability distribution function but has the unusual properties

$$\int_{-\infty}^{\infty} \delta(x)dx = 1 \tag{1.27}$$

$$\int_{-\infty}^{\infty} g(x)\delta(x)dx = g(0), \tag{1.28}$$

where $g(x)$ is any reasonable continuous function. It can loosely be thought of as a function with properties

$$\delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & x \neq 0 \end{cases} . \tag{1.29}$$

Example: Consider the exponential function

$$\mathrm{f}(x) = \begin{cases} 0, & x < 0 \\ a\exp[-ax], & x \ge 0 \end{cases} \tag{1.30}$$

with $a > 0$. This function satisfies the requirements for it to be a probability distribution function:

- It is single valued.
- It is greater than or equal to 0 everywhere.
- It is normalized:

$$\int_{-\infty}^{\infty} \mathrm{f}(x)\,dx = \int_{0}^{\infty} a\exp[-ax]\,dx = -\exp[-ax]\Big|_{0}^{\infty} = 1.$$

- Its integral exists for all intervals:

$$P(x_a \le x \le x_b) = \int_{x_a}^{x_b} \mathrm{f}(x)\,dx = \exp[-x_a] - \exp[-x_b] \quad \text{for } 0 \le x_a \le x_b,$$

with analogous results if the limits are less than 0.

Continued on page 8

The exponential probability distribution is widely applicable. For example, the distribution of power in the power spectrum of white noise and the distribution of intervals between decays of a radioactive source are given by an exponential probability distribution.

1.2.2 Cumulative Probability Distribution Function

The cumulative distribution function, $F(x)$, is defined to be

$$F(x) = \int_{-\infty}^{x} \mathrm{f}(y)\, dy. \tag{1.31}$$

The relation between $\mathrm{f}(x)$ and $F(x)$ is shown graphically in Figure 1.4. Since $F(x)$ is the integral over a probability density function, it is a true probability—the probability that x lies between $-\infty$ and x. The cumulative distribution function for the Dirac delta function is a step function, sometimes called the *Heaviside function*:

$$H(x) = \int_{-\infty}^{x} \delta(y)\, dy = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases}. \tag{1.32}$$

In fact, the delta function can be defined as the derivative of the Heaviside function. The cumulative distribution function is useful for working with sparse or noisy data.

1.2.3 Change of Variables

Suppose we wish to change the independent variable in a probability distribution function $\mathrm{f}(x)$ to a different variable y, where the coordinate transformation is given by $x(y)$. We require that the probability in an interval dx be the same as the probability in the

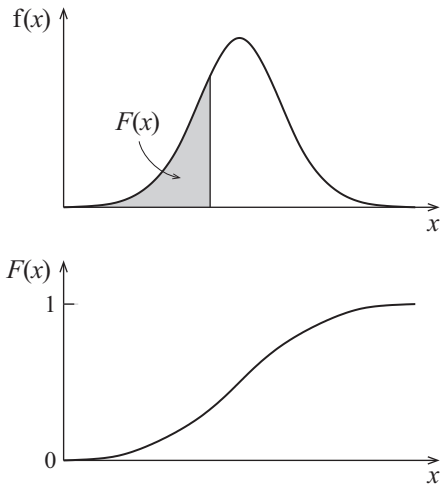

Figure 1.4: The relation between the probability distribution function f(x) and the cumulative distribution function $F(x)$ derived from it.

corresponding dy. The probability distribution for y is then given by

$$\mathrm{f}(x)dx = \mathrm{f}(x(y))\left|\frac{dx}{dy}\right|dy = \mathrm{g}(y)dy, \tag{1.33}$$

from which we find

$$\mathrm{g}(y) = \mathrm{f}(x(y))\left|\frac{dx}{dy}\right|. \tag{1.34}$$

The factor $|dx/dy|$ accounts for the difference between the lengths of dy and dx.

1.3 Characterizations of Probability Distributions

The full description of a probability distribution is the entire distribution itself, given either by its functional form or by an equivalent, such as a table, a graph, or the complete set of moments of the distribution. It is often more convenient to deal with a small set of quantities that succinctly describe the most important properties of a distribution. The descriptors could be a single number that summarizes the entire distribution, such as the mean, mode, or median; a measure of the width of the distribution, such as its variance or full width at half maximum; a measure of the asymmetry, such as skewness; or, more generally, a few of the lower moments of the distribution.

1.3.1 Medians, Modes, and Full Width at Half Maximum

A simple way to summarize a probability distribution $\mathrm{f}(x)$ with a single number is to give the value of x at which the distribution reaches its maximum value. This is called the mode of a distribution, x_{mode}. For a continuous function with a single peak, the mode can be calculated from

$$\left.\frac{d\,\mathrm{f}(x)}{dx}\right|_{x_{mode}} = 0. \tag{1.35}$$

The mode is most useful for probability distributions with a single maximum or just one dominant maximum.

Another useful single-number descriptor is the median x_{median}, defined by

$$\frac{1}{2} = \int_{-\infty}^{x_{median}} \mathrm{f}(x)\,dx. \tag{1.36}$$

The median is the "middle" of a distribution in the sense that a sample from $\mathrm{f}(x)$ has equal probabilities of lying above and below x_{median}. The median is particularly useful when one wants to reduce the influence of distant outliers or long tails of a distribution.

The full width at half maximum (often abbreviated by FWHM) is a useful and easily measured descriptor of the width of a probability distribution. The half maximum of a distribution is $\mathrm{f}(x_{mode})/2$. To calculate the FWHM, find the values a and b such that

$$\mathrm{f}(a) = \mathrm{f}(b) = \frac{1}{2}\mathrm{f}(x_{mode}); \tag{1.37}$$

then FWHM $= b - a$. Like the mode from which it is derived, the FWHM is most useful for probability distributions with a single maximum or just one dominant maximum; it may not be a sensible way to describe more complicated distributions.

1.3.2 Moments, Means, and Variances

The mth moment M_m of a continuous probability distribution function $\mathrm{f}(x)$ is defined to be

$$M_m = \int_{-\infty}^{\infty} x^m \, \mathrm{f}(x)\, dx. \tag{1.38}$$

If $P(A_j)$ is a discrete probability distribution and the outcomes A_j are real numbers, the moments of $P(A_j)$ are

$$M_m = \sum_j A_j^m \, P(A_j), \tag{1.39}$$

where the sum is taken over all possible values of j. Moment M_m is called the mean value of x^m, symbolically, $M_m = \langle x^m \rangle$. It can be thought of as the average of x^m weighted by the probability that x will occur.

The zeroth moment is equal to 1 for a correctly normalized distribution function:

$$M_0 = \int_{-\infty}^{\infty} x^0 \, \mathrm{f}(x)\, dx = \int_{-\infty}^{\infty} \mathrm{f}(x)\, dx = 1. \tag{1.40}$$

The first moment is

$$M_1 = \langle x \rangle = \int_{-\infty}^{\infty} x \, \mathrm{f}(x)\, dx \tag{1.41}$$

or, for a discrete probability distribution,

$$M_1 = \langle A \rangle = \sum_j A_j \, P(A_j). \tag{1.42}$$

The first moment is usually called the *mean*. To gain a feel for the significance of the mean, suppose that n samples a_i have been generated from the discrete probability distribution $P(A_j)$ and that the number of times each value A_j has been generated is k_j. By definition

$$P(A_j) = \lim_{n \to \infty} \frac{k_j}{n}, \tag{1.43}$$

so the mean is

$$\langle A \rangle = \lim_{n \to \infty} \frac{1}{n} \sum_j A_j k_j. \tag{1.44}$$

Since the quantity k_j means that A_j appears k_j times, we can list all the a_i individually, producing a list of n values of a_i, $i = 1, \ldots, n$, with each value A_j appearing k_j times in the list. The weighted sum in equation 1.44 can therefore be replaced by the unweighted sum over all the individual values of a_i, and the mean value becomes

$$\langle x \rangle = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} a_i. \tag{1.45}$$

This corresponds to our intuitive understanding of a mean value.

The mean is yet another way to describe a probability distribution function with a single number. If $\mathrm{f}(x)$ is symmetric about x_{median}, then $M_1 = x_{median}$. There is no guarantee that

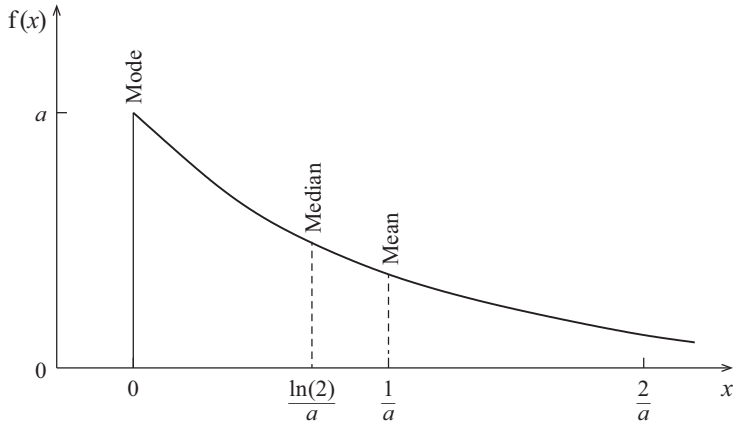

Figure 1.5: The mean, median, and mode of the exponential distribution function $f(x) = a\exp[-ax]$.

a probability distribution has a mean, though. The *Lorentzian distribution* (also called the *Cauchy distribution*)

$$f(x) = \frac{1}{\pi}\frac{b}{b^2+(x-a)^2} \tag{1.46}$$

satisfies equations 1.23–1.25 and is a perfectly valid probability distribution function, but it does not have a mean. As x becomes large, $xf(x)$ becomes proportional to $1/x$, and the integral of $xf(x)$ approaches a logarithmic function, which increases without limit. The integral in equation 1.41 is therefore undefined, and the mean does not exist.[1] The mode and median of the Lorentzian function do exist and are $x_{mode} = x_{median} = a$. They provide alternative single-number descriptions of the Lorentzian distribution.

Example: Find the mean, median, and mode of the exponential probability distribution function

$$f(x) = a\exp[-ax], \qquad x \geq 0.$$

By inspection, the maximum of the function occurs at $x = 0$, so the mode is $x_{mode} = 0$. The median is given by

$$\frac{1}{2} = \int_{-\infty}^{x_{median}} f(x)dx = \int_0^{x_{median}} a\exp[-ax]\,dx = 1-\exp[-ax_{median}],$$

which yields

$$x_{median} = \frac{1}{a}\ln 2.$$

The mean value of x is

$$\langle x\rangle = \int_{-\infty}^{\infty} x\,a\exp[-ax]\,dx = -\frac{1}{a}(1+x)\exp[-x]\Big|_0^{\infty} = \frac{1}{a}.$$

The mean, median, and mode for the exponential function are shown in Figure 1.5.

[1] If equation 1.5 is taken to be an improper integral, $\langle x\rangle$ does exist for a Lorentzian distribution function, and $\langle x\rangle = a$. The higher moments of the distribution remain undefined, though.

The concept of mean value can be generalized to the mean value of a function. Suppose that A has possible values A_j and that the probability distribution function for the A_j is $P(A_j)$. Also suppose that $g(A)$ is a function of A. The mean value of g is

$$\langle g \rangle = \sum_j g(A_j) P(A_j). \tag{1.47}$$

If $\mathrm{f}(x)$ is a continuous probability distribution function and $g(x)$ is also a continuous function, the mean value of $g(x)$, denoted by $\langle g(x) \rangle$, is

$$\langle g(x) \rangle = \int_{-\infty}^{\infty} g(x) \mathrm{f}(x)\, dx. \tag{1.48}$$

The quantity $\langle g(x) \rangle$ is a weighted average of $g(x)$, where the weight is the probability that x occurs. For example, $\mathrm{f}(x)$ might be the probability distribution for the radii of raindrops in a storm and $g(x)$ the mass of the raindrop as a function of radius. Then $\langle g(x) \rangle$ is the average mass of raindrops. For completeness we note that

$$\langle a_1 g_1(x) + a_2 g_2(x) \rangle = \int_{-\infty}^{\infty} [a_1 g_1(x) + a_2 g_2(x)] \mathrm{f}(x)\, dx = a_1 \langle g_1(x) \rangle + a_2 \langle g_2(x) \rangle, \tag{1.49}$$

so calculating a mean is a linear operation.

The *variance*, σ^2, of a distribution is defined to be the mean value of $(x - \langle x \rangle)^2$:

$$\sigma^2 = \langle (x - \mu)^2 \rangle, \tag{1.50}$$

where for notational convenience we have set $\langle x \rangle = \mu$. The positive square root of the variance, denoted by σ, is called the *standard deviation*. The variance and the standard deviation are measures of the width or spread of a distribution function. Their precise meaning depends on the specific distribution, but roughly half of the probability lies between $\mu - \sigma$ and $\mu + \sigma$. The variance of a distribution is related to the second moment of the distribution. Expanding the expression for σ^2, we find

$$\begin{aligned} \sigma^2 = \langle x^2 - 2x\mu + \mu^2 \rangle &= \langle x^2 \rangle - 2\mu \langle x \rangle + \mu^2 \\ &= M_2 - \mu^2 \qquad (1.51) \\ &= M_2 - M_1^2. \qquad (1.52) \end{aligned}$$

The variance does not exist for all functions—it does not exist for the Lorentzian function because M_2 diverges. But the variance is not the only way to characterize the width of a distribution function, and even when it exists, it is not necessarily the best way to characterize the width. For example, the quantity $\langle |x - \mu| \rangle$ also measures the width and since it is linear in $x - \mu$, not quadratic, it is less sensitive to parts of the distribution far from μ than is the variance. Or one may want to use the FWHM, which does exist for the Lorentzian. The variance may not even be the most informative way to describe the width of a distribution. The variance of a rectangular distribution function is less useful than the full width of the rectangle.

The asymmetry of a probability distribution can be characterized by its *skewness*. There are three common definitions of skewness:

$$\text{skewness} = \frac{\text{mean} - \text{mode}}{\text{standard deviation}} = \frac{\langle x \rangle - x_{mode}}{\sigma}, \tag{1.53}$$

$$\text{skewness} = \frac{\text{mean} - \text{median}}{\text{standard deviation}} = \frac{\langle x \rangle - x_{median}}{\sigma}, \tag{1.54}$$

$$\text{skewness} = \frac{\langle (x-\mu)^3 \rangle}{\sigma^3}. \tag{1.55}$$

One must always specify which definition is being used.

Example: Calculate the second moment, the variance, and the skewness of the exponential distribution function

$$\mathrm{f}(x) = a \exp[-ax], \qquad x \geq 0.$$

The second moment is

$$\begin{aligned} M_2 &= \int_{-\infty}^{\infty} x^2 \, \mathrm{f}(x) \, dx \\ &= \int_0^{\infty} ax^2 \exp[-ax] \, dx = -\frac{1}{a^2}[(ax)^2 + 2ax + 2] \exp[-ax] \Big|_0^{\infty} \\ &= \frac{2}{a^2}. \end{aligned}$$

Calculate σ from equation 1.52 and the value of μ just found:

$$\sigma^2 = M_2 - \mu^2 = \frac{2}{a^2} - \left(\frac{1}{a}\right)^2 = \frac{1}{a^2}.$$

The standard deviation is shown in Figure 1.6. For comparison, the probability distribution drops to half its maximum value at $x = 0$ and at

$$a \exp[-ax_{1/2}] = \frac{1}{2},$$

so its full width at half maximum is

$$\text{FWHM} = \frac{1}{a} \ln(2a).$$

We have already shown that $\langle x \rangle = \mu = 1/a$ and $x_{median} = \ln(2)/a$ for the exponential distribution. Then the skewness as defined by equation 1.54 is

$$\text{skewness} = \frac{\langle x \rangle - x_{median}}{\sigma} = \frac{1/a - \ln(2)/a}{1/a^2} = a(1 - \ln 2).$$

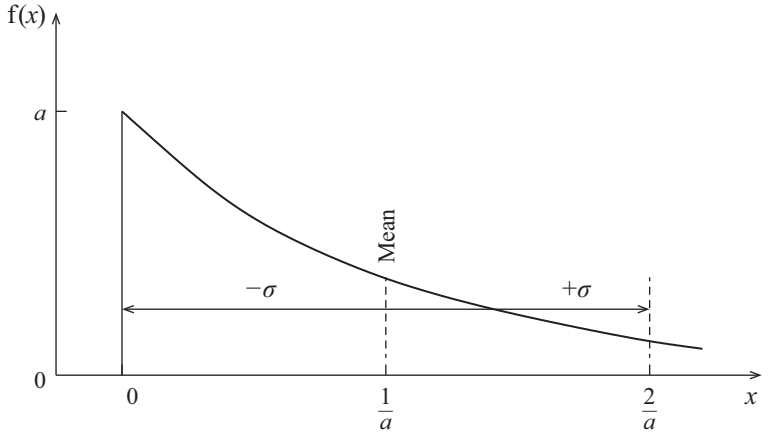

Figure 1.6: The mean $< x >$ and standard deviation σ of the exponential distribution function $f(x) = a\exp[-ax]$.

1.3.3 Moment Generating Function and the Characteristic Function

The moment generating function provides a convenient way to calculate moments for some probability distributions. For a continuous probability distribution $f(x)$, the moment generating function is defined to be

$$M(\zeta) = \int_{-\infty}^{\infty} \exp[\zeta x] f(x) dx, \tag{1.56}$$

where the integral needs to exist only in the neighborhood of $\zeta = 0$. To understand the moment generating function, expand $\exp[\zeta x]$ in a Taylor series and carry out the integration term by term. The moment generating function becomes

$$M(\zeta) = \int_{-\infty}^{\infty} (1 + \zeta x + \frac{\zeta^2 x^2}{2!} + \cdots) f(x) dx$$

$$= M_0 + \zeta M_1 + \frac{\zeta^2}{2!} M_2 + \cdots . \tag{1.57}$$

For a correctly normalized probability distribution, $M_0 = M(0) = 1$. The higher moments of $f(x)$ can be calculated from the derivatives of the moment generating function:

$$M_m = \left. \frac{\partial^m M(\zeta)}{\partial \zeta^m} \right|_{\zeta=0}, \qquad m \geq 1. \tag{1.58}$$

Example: The moment generating function for the exponential probability distribution

$$f(x) = a\exp[-ax], \quad x \geq 0,$$

is

$$M(\zeta) = a\int_0^\infty \exp[\zeta x]\exp[-ax]\,dx = a\int_0^\infty \exp[(\zeta - a)x]\,dx = \frac{a}{a-\zeta}.$$

The first and second moments of the probability distribution are

$$M_1 = \left.\frac{\partial M(\zeta)}{\partial \zeta}\right|_{\zeta=0} = \left.\frac{a}{(a-\zeta)^2}\right|_{\zeta=0} = \frac{1}{a}$$

$$M_2 = \left.\frac{\partial^2 M(\zeta)}{\partial \zeta^2}\right|_{\zeta=0} = \left.\frac{2a}{(a-\zeta)^3}\right|_{\zeta=0} = \frac{2}{a^2}.$$

The *characteristic function* $\phi(\nu)$ of a continuous probability distribution function f(x) is defined to be

$$\phi(\nu) = \int_{-\infty}^{\infty} \exp[i\nu x]\,\mathrm{f}(x)dx, \tag{1.59}$$

where $i = \sqrt{-1}$, and ν is a real number. Looking ahead to Chapter 8 on Fourier analysis, we recognize the characteristic function as the inverse Fourier transform of f(x) (see equation 8.61). Expanding the exponential, we have

$$\begin{aligned}\phi(\nu) &= \int_{-\infty}^{\infty}\left(1 + i\nu x + \frac{1}{2!}(i\nu)^2x^2 + \frac{1}{3!}(i\nu)^3x^3 + \cdots\right)\mathrm{f}(x)dx \\ &= 1 + i\nu M_1 + \frac{(i\nu)^2}{2!}M_2 + \cdots + \frac{(i\nu)^n}{n!}M_n + \cdots.\end{aligned} \tag{1.60}$$

The Fourier transform of $\phi(\nu)$ is

$$\begin{aligned}\frac{1}{2\pi}\int_\nu \phi(\nu)\exp[-i\nu x']d\nu &= \frac{1}{2\pi}\int_\nu\left\{\int_x \exp[i\nu x]\,\mathrm{f}(x)dx\right\}\exp[-i\nu x']d\nu \\ &= \int_x \mathrm{f}(x)\left\{\frac{1}{2\pi}\int_\nu \exp[i\nu(x-x')]d\nu\right\}dx \\ &= \int_x \mathrm{f}(x)\,\delta(x-x')\,dx \\ &= \mathrm{f}(x'),\end{aligned} \tag{1.61}$$

where $\delta(x-x')$ is the Dirac delta function (see Table 8.1,which lists some Fourier transform pairs). As expected, the original probability distribution function is the Fourier transform of $\phi(\nu)$.

Combining equations 1.60 and 1.61 we find

$$\mathrm{f}(x') = \frac{1}{2\pi}\int_\nu\left[1 + i\nu M_1 + \frac{(i\nu)^2}{2!}M_2 + \cdots + \frac{(i\nu)^n}{n!}M_n + \cdots\right]\exp[-i\nu x']d\nu. \tag{1.62}$$

This shows that a probability distribution function can be reconstructed from the values of its moments, a remarkable and unexpected relation between the Fourier transform and

the moment equation (equation 1.38). We will need this result when deriving the Gaussian probability distribution by way of the central limit theorem.

1.4 Multivariate Probability Distributions

1.4.1 Distributions with Two Independent Variables

Probability distribution functions can have more than one independent variable (see Figure 1.7). A function of two independent variables x_1 and x_2 is a valid probability distribution function if it is single-valued and satisfies the requirements:

$$\mathrm{f}(x_1,x_2) \geq 0 \tag{1.63}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{f}(x_1,x_2)\,dx_1 dx_2 = 1 \tag{1.64}$$

$$P(a_1 < x_1 < b_1,\, a_2 < x_2 < b_2) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} \mathrm{f}(x_1,x_2)\,dx_1 dx_2, \tag{1.65}$$

where $P(a_1 < x_1 < b_1,\, a_2 < x_2 < b_2)$ is the probability that x_1 lies in the range $a_1 < x_1 < b_1$ and x_2 lies in the range $a_2 < x_2 < b_2$. Equation 1.65 can be generalized to the probability that x_1 and x_2 lie in an arbitrary area A in the (x_1,x_2) plane:

$$P(x_1,x_2 \subset A) \;=\; \int_A \mathrm{f}(x_1,x_2)\,dx_1 dx_2. \tag{1.66}$$

There are two *marginal probability distributions*:

$$\mathrm{g}_1(x_1) = \int_{x_2=-\infty}^{\infty} \mathrm{f}(x_1,x_2)\,dx_2 \tag{1.67}$$

$$\mathrm{g}_2(x_2) = \int_{x_1=-\infty}^{\infty} \mathrm{f}(x_1,x_2)\,dx_1. \tag{1.68}$$

Thus, $\mathrm{g}_1(x_1)$ is the integral of all the probability that is spread out in x_2 (see Figure 1.7).

Given a two-parameter distribution $\mathrm{f}(x_1,x_2)$ and its marginal probability $\mathrm{g}_1(x_1)$, the *conditional distribution* $\mathrm{h}(x_2|x_1)$ is defined to be

$$\mathrm{h}(x_2|x_1) \;=\; \frac{\mathrm{f}(x_1,x_2)}{\mathrm{g}_1(x_1)}. \tag{1.69}$$

The conditional distribution is a cut through $\mathrm{f}(x_1,x_2)$ at some constant value of x_1. Compare this to the marginal distribution, which is an integral, not a cut. The conditional distribution is already properly normalized, since

$$\int_{-\infty}^{\infty} \mathrm{h}(x_2|x_1)dx_2 \;=\; \int_{-\infty}^{\infty} \frac{\mathrm{f}(x_1,x_2)}{\mathrm{g}_1(x_1)}\,dx_2 \;=\; \frac{1}{\mathrm{g}_1(x_1)}\int_{-\infty}^{\infty} \mathrm{f}(x_1,x_2)\,dx_2 \;=\; \frac{\mathrm{g}_1(x_1)}{\mathrm{g}_1(x_1)} \;=\; 1. \tag{1.70}$$

We say that x_2 is independent of x_1 if the probability that x_2 occurs is independent of the probability that x_1 occurs. The conditional probability then becomes

$$\mathrm{h}(x_2|x_1) \;=\; \mathrm{h}(x_2). \tag{1.71}$$

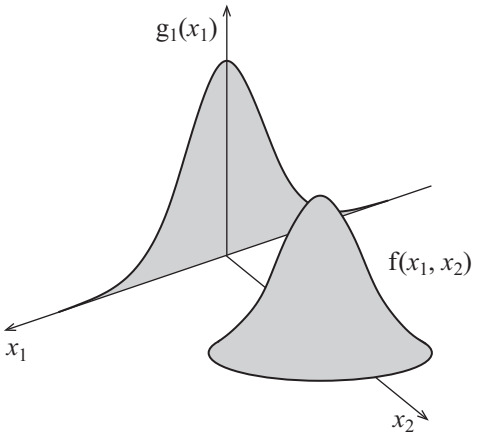

Figure 1.7: A two-dimensional distribution function, $f(x_1,x_2)$, and the marginal distribution, $g_1(x_1)$, derived from it.

If x_1 and x_2 are independent, equation 1.69 can be rewritten as

$$f(x_1,x_2) = g_1(x_1)h(x_2|x_1) = g_1(x_1)h(x_2), \tag{1.72}$$

so the probability that both x_1 and x_2 will occur is the product of the individual probabilities of their occurrence.

Moments can be calculated for each of the variables individually:

$$\langle x_1^m \rangle = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} x_1^m f(x_1,x_2) dx_1 dx_2 \tag{1.73}$$

$$\langle x_2^n \rangle = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} x_2^n f(x_1,x_2) dx_1 dx_2, \tag{1.74}$$

but it is also possible to calculate joint moments:

$$\langle x_1^m x_2^n \rangle = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} x_1^m x_2^n f(x_1,x_2) dx_1 dx_2. \tag{1.75}$$

If x_1 and x_2 are independent, the integrals over x_1 and x_2 in equation 1.75 separate, and $\langle x_1^m x_2^n \rangle = \langle x_1^m \rangle \langle x_2^n \rangle$.

1.4.2 Covariance

If a probability distribution function has two independent variables x_1 and x_2, the *covariance* between x_1 and x_2 is defined to be

$$\sigma_{12} = \sigma_{21} = \langle (x_1 - \langle x_1 \rangle)(x_2 - \langle x_2 \rangle) \rangle \tag{1.76}$$

$$= \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} (x_1 - \langle x_1 \rangle)(x_2 - \langle x_2 \rangle) f(x_1,x_2) dx_1 dx_2. \tag{1.77}$$

The covariance measures the extent to which the value of a sample from x_1 depends on the value of a sample from x_2. If x_1 is independent of x_2, then $f(x_1,x_2) = f_1(x_1)f_2(x_2)$, and the

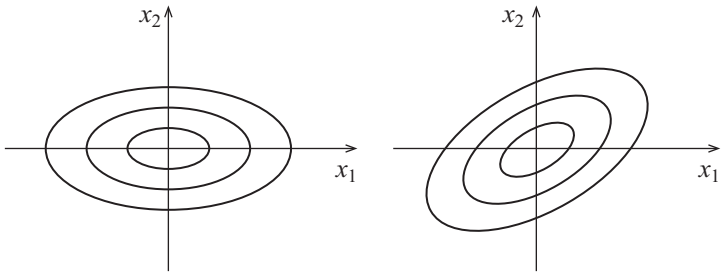

Figure 1.8: Contour plots of two two-parameter probability distributions $f(x_1,x_2)$ for which the contours of constant probability are ellipses centered on the origin. The covariance between x_1 and x_2 is zero for the distribution on the left, but it is nonzero for the distribution on the right.

covariance becomes

$$\begin{aligned}\sigma_{12} &= \int_{x_1}\int_{x_2}(x_1-\langle x_1\rangle)(x_2-\langle x_2\rangle)f_1(x_1)f_2(x_2)dx_1dx_2\\ &= \left[\int(x_1-\langle x_1\rangle)f_1(x_1)dx_1\right]\left[\int(x_2-\langle x_2\rangle)f_2(x_2)dx_2\right]\\ &= (\langle x_1\rangle-\langle x_1\rangle)\,(\langle x_2\rangle-\langle x_2\rangle)\\ &= 0.\end{aligned} \tag{1.78}$$

To gain some insight into the meaning of covariance, consider a two-parameter distribution for which the contours of constant probability are ellipses centered on the origin. Two possible contour plots of the distribution are shown in Figure 1.8. Ellipses centered on the origin have the general functional form $a_1x_1^2+a_{12}x_1x_2+a_2x_2^2=$ constant, so the probability distribution must have the form $f(a_1x_1^2+a_{12}x_1x_2+a_2x_2^2)$. If $a_{12}=0$, the major and minor axes of the ellipses are horizontal and vertical, and they coincide with the coordinate axes as shown in the left panel of the figure. If $a_{12}>0$, the axes are tilted as shown in the right panel. Consider the case

$$f(x_1,x_2) \;=\; a_1x_1^2+a_{12}x_1x_2+a_2x_2^2 \tag{1.79}$$

with $-\alpha\le x_1\le\alpha$ and $-\beta\le x_2\le\beta$. Since the ellipses are symmetric about the origin, the mean values of x_1 and x_2 are zero. The covariance becomes

$$\begin{aligned}\sigma_{12} &= \int_{x_1=-\alpha}^{\alpha}\int_{x_2=-\beta}^{\beta}x_1x_2(a_1x_1^2+a_{12}x_1x_2+a_2x_2^2)dx_1dx_2\\ &= \int_{x_1=-\alpha}^{\alpha}\int_{x_2=-\beta}^{\beta}(a_1x_1^3x_2+a_{12}x_1^2x_2^2+a_2x_1x_2^3)dx_1dx_2\\ &= \frac{4a_{12}}{9}\alpha^3\beta^3.\end{aligned} \tag{1.80}$$

The covariance is nonzero if a_{12} is nonzero, so the covariance is nonzero if the ellipses are tilted.

1.4.3 Distributions with Many Independent Variables

These results for two independent variables generalize easily to many independent variables. The multivariate function $\mathrm{f}(x_1, x_2, \ldots, x_n)$ can be a probability distribution function if it is single valued and if

- $$\mathrm{f}(x_1, x_2, \ldots, x_n) \geq 0 \tag{1.81}$$
- $$\int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} \cdots \int_{x_n=-\infty}^{\infty} \mathrm{f}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n = 1 \tag{1.82}$$
- $$P(a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_n < x_n < b_n) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} \mathrm{f}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n, \tag{1.83}$$

where P is the probability that x_1 lies in the range $a_1 < x_1 < b_1$, that x_2 lies in the range $a_2 < x_2 < b_2$, and so on. If V is an arbitrary volume in the n-dimensional space spanned by $(x_1, x_2, \ldots, x_n)$, the probability that a point lies within the volume is

$$P(x_1, x_2, \ldots, x_n \subset V) = \int_V \mathrm{f}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n. \tag{1.84}$$

Marginal and Conditional Distributions: Many different marginal probability distributions can be constructed from $\mathrm{f}(x_1, x_2, \ldots, x_n)$. For example, there is a marginal distribution given by integrating over x_3 and x_5,

$$\mathrm{g}(x_1, x_2, x_4, x_6, \ldots, x_n) = \int_{x_3=-\infty}^{\infty} \int_{x_5=-\infty}^{\infty} \mathrm{f}(x_1, x_2, \ldots, x_n)\, dx_3 dx_5; \tag{1.85}$$

and there is a conditional distribution for each marginal distribution. For example, the conditional distribution $\mathrm{h}(x_3, x_5 | x_1, x_2, x_4, x_6, \ldots, x_n)$ is given by

$$\mathrm{h}(x_3, x_5 | x_1, x_2, x_4, x_6, \ldots, x_n) = \frac{\mathrm{f}(x_1, x_2, \ldots, x_n)}{\mathrm{g}(x_1, x_2, x_4, x_6, \ldots, x_n)}. \tag{1.86}$$

In fact, $\mathrm{f}(x_1, x_2, \ldots, x_n)$ can be marginalized in any direction, not just along the $(x_1, x_2, \ldots, x_n)$ coordinate axes. To marginalize in any other desired direction, simply rotate the coordinate system so that one of the rotated coordinate axes points in the desired direction. Then integrate the distribution along that coordinate.

If variables x_i and x_j are independent of each other and their probability distribution functions are $\mathrm{f}_i(x_i)$ and $\mathrm{f}_j(x_j)$, the joint probability distribution function can be written

$$\mathrm{f}(x_1, x_2, \ldots, x_n) = \mathrm{f}_i(x_i)\mathrm{f}_j(x_j)\bar{\mathrm{f}}_{ij}, \tag{1.87}$$

where $\bar{\mathrm{f}}_{ij}$ is short-hand notation for the probability distribution function of all the other variables. If all the x_i are all independent of one another, their joint probability distribution function is the product of their individual probability distribution functions:

$$\mathrm{f}(x_1, x_2, \ldots, x_n) = \mathrm{f}_1(x_1)\mathrm{f}_2(x_2) \cdots \mathrm{f}_n(x_n). \tag{1.88}$$

Covariances: The covariance between parameters x_i and x_j, defined to be

$$\sigma_{ij} = \sigma_{ji} = \left\langle \left(x_i - \langle x_i \rangle\right)\left(x_j - \langle x_j \rangle\right)\right\rangle \tag{1.89}$$

$$= \int \cdots \int \left(x_i - \langle x_i \rangle\right)\left(x_j - \langle x_j \rangle\right) \mathrm{f}(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n, \tag{1.90}$$

measures the extent to which the value of x_i depends on the value of x_j. If $i = j$, equation 1.90 becomes

$$\sigma_{ii} = \left\langle \left(x_i - \langle x_i \rangle\right)^2 \right\rangle, \tag{1.91}$$

which is the variance of x_i, so it is common to use the notation $\sigma_{ii} = \sigma_i^2$.

If any pair of parameters x_i and x_j are independent of each other, then, using the notation of equation 1.87, their covariance is

$$\begin{aligned}
\sigma_{ij} &= \int \cdots \int \left(x_i - \langle x_i \rangle\right)\left(x_j - \langle x_j \rangle\right) \mathrm{f}_i(x_i)\mathrm{f}_j(x_j)\bar{\mathrm{f}}_{ij}\, dx_n \\
&= \left[\int \left(x_i - \langle x_i \rangle\right)\mathrm{f}_i(x_i) dx_i\right]\left[\int \left(x_j - \langle x_j \rangle\right)\mathrm{f}_j(x_j) dx_j\right] \\
&= \left(\langle x_i \rangle - \langle x_i \rangle\right)\left(\langle x_j \rangle - \langle x_j \rangle\right) \\
&= 0.
\end{aligned} \tag{1.92}$$

Moment Generating Function: The moment generating function for a multivariate probability distribution is

$$M(\zeta_1, \zeta_2, \ldots, \zeta_n) = \int_{x_1=-\infty}^{\infty} \cdots \int_{x_n=-\infty}^{\infty} \exp[\zeta_1 x_1 + \zeta_2 x_2 + \cdots + \zeta_n x_n] \times \mathrm{f}(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n. \tag{1.93}$$

The first and second moments of the probability distribution can be calculated by taking first and second derivatives, respectively, of the moment generating function:

$$\langle x_j \rangle = \left.\frac{\partial M(\zeta_1, \zeta_2, \ldots, \zeta_n)}{\partial \zeta_j}\right|_{\zeta_1 = \zeta_2 = \cdots = \zeta_n = 0} \tag{1.94}$$

$$\langle x_j x_k \rangle = \left.\frac{\partial^2 M(\zeta_1, \zeta_2, \ldots, \zeta_n)}{\partial \zeta_j \partial \zeta_k}\right|_{\zeta_1 = \zeta_2 = \cdots = \zeta_n = 0} \tag{1.95}$$

For those who collect inscrutable formulas, the general case is

$$\left\langle x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n} \right\rangle = \left.\frac{\partial^{m_1 + m_2 + \cdots + m_n} M(\zeta_1, \zeta_2, \ldots, \zeta_n)}{\partial \zeta_1^{m_1} \partial \zeta_2^{m_2} \cdots \partial \zeta_n^{m_n}}\right|_{\zeta_1 = \zeta_2 = \cdots = \zeta_n = 0}. \tag{1.96}$$

Transformation of Variables: Finally, it is sometimes necessary to transform a probability distribution function $\mathrm{f}(x_1, x_2, \ldots, x_n)$ to a different set of variables $(y_1, y_2, \ldots, y_n)$. Let the equations for the coordinate transformation be

$$\begin{aligned}
x_1 &= x_1(y_1, y_2, \ldots, y_n) \\
&\vdots \quad \vdots \\
x_n &= x_n(y_1, y_2, \ldots, y_n).
\end{aligned} \tag{1.97}$$

The transformation must be invertible for the probability distribution in the new coordinates to be meaningful. Following standard methods for changing variables in multiple integrals,[2] we have

$$\begin{aligned} \mathrm{f}(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n &= \mathrm{f}(y_1, y_2, \ldots, y_n) \left| \frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)} \right| dy_1 dy_2 \cdots dy_n \\ &= g(y_1, y_2, \ldots, y_n) dy_1 dy_2 \cdots dy_n, \end{aligned} \tag{1.98}$$

from which we find

$$g(y_1, y_2, \ldots, y_n) = \mathrm{f}(y_1, y_2, \ldots, y_n) \left| \frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)} \right|. \tag{1.99}$$

The Jacobian determinant in equation 1.99 accounts for the difference between the sizes of the volume elements in the two coordinate systems.

[2] See, for example, Riley et al. (2002).