Chapter 1
The Theory of Probability

1.1 STATISTICAL MECHANICS AND PROBABILITY

Statistical mechanics is the theoretical bridge between the microscopic world and the macroscopic world. The microscopic world is a quantum mechanical world, in which electrons, protons, and neutrons organize themselves into atoms, molecules, crystals, and other structures on an atomic length scale. The macroscopic world is the tangible world of trees and rocks and water. The microscopic structure of these things is usually completely hidden from our view. It is only by a complex chain of reasoning that the fundamental laws of physics, which govern the structure and interactions of the unseen microscopic constituents of things, can be shown to give rise to the observed physical properties of ordinary objects. That chain of reasoning is the subject of statistical mechanics.

A typical macroscopic object, a glass of water, contains about 3×10^{25} molecules. The extreme compactness of exponential notation masks the vastness of that number. If that glass of water were mixed throughout all the oceans of the world, then a glass of water, drawn anywhere on earth, would contain many of the molecules that had been in it. Clearly, there is no possibility of ever knowing the exact microscopic state of a glass of water. In statistical mechanics we will have to work with very incomplete information about the systems we are analyzing. Typically, we can express our information about the state of the system in 3 or 4 numbers, such as the system's volume, mass, temperature, and pressure, while a complete specification would need about 10^{26} separate numbers. It is remarkable that, with this meager information, we can predict so many important things with high precision and reliability. For other things we can make only probabilistic predictions. But, even for those, the methods of statistical mechanics permit us to make accurate predictions of the probabilities.

The general mathematical discipline involved in making predictions with limited information is probability theory. In a real sense, statistical mechanics is a branch of probability theory. However, the systems under analysis in statistical mechanics have certain special characteristics that have allowed the subject to develop a subtlety and predictive power that are unmatched by any other subject in which statistical methods are used. This chapter will be devoted to an introduction to the general ideas of probability theory. All the following chapters will specialize the theory to the problem of calculating the physical properties of aggregate matter.

1.2 THE DEFINITION OF PROBABILITY

The reader certainly has some idea of what probability means and most likely knows how to make simple calculations of probabilities. In order to construct a mathematical theory of probability, we will first look at a problem in probability that is easy to solve and then abstract the rules we use to solve that problem and restate them in a more precise and general way.

The problem we will consider is to calculate the probability of throwing six at dice. It will help to

assume that the two dice are of different colors, let us say red and blue. Each time the dice are thrown we get two numbers. The number on the red die we call R, and that on the blue we call B. The pair of numbers (R, B) will be denoted by the single symbol x. The value of x is always some element from the 36-element set

$$\Omega = \{(1,1), (1,2), \ldots, (6,5), (6,6)\} \tag{1.1}$$

If the dice are well made, it is reasonable to assume that each of the 36 possible values of x has the same probability of occurrence. The probability of occurrence of x is defined as the fraction of times that x occurs in a very long sequence of throws. We write it as $P(x)$. Regardless of whether the dice are well made or not, two things follow immediately from the definition of $P(x)$. They are that

$$0 \leq P(x) \leq 1 \tag{1.2}$$

and

$$\sum_{x \in \Omega} P(x) = 1 \tag{1.3}$$

For well-made dice, $P(x) = 1/36$ for all x. To calculate the probability of throwing six we must first see which combinations give $R + B = 6$. They are

$$x = (1,5), (2,4), (3,3), (4,2), \text{ and } (5,1) \tag{1.4}$$

We then sum $P(x)$ over those five values of x to get the probability that $R + B = 6$.

$$\mathbf{P}[R{+}B{=}6] = \frac{5}{36} \tag{1.5}$$

Now let us try to generalize the ideas used in this dice problem.

We assume that we are given a system (for example, the dice) whose possible configurations or states are described by a variable x that can take values within some set Ω. The elements of Ω we will call the possible *microstates* of the system and Ω will be called the *microstate space*.* At this point it is assumed that the set Ω contains only a finite number of microstates. Later, infinite and continuous sets of microstates will be considered.

There exists some procedure for constructing a sampling sequence of arbitrary length, x_1, x_2, ..., where $x_n \in \Omega$. (For example, repeatedly shake the dice and throw them from a height of at least 2 feet.) The sampling sequence must have the property that the function

$$P(x) \equiv \lim_{N \to \infty} \Big(\frac{\text{The number of times } x_n = x \text{ for } n = 1, \ldots, N}{N} \Big) \tag{1.6}$$

exists for all $x \in \Omega$. This definition of the probability function, $P(x)$, in terms of a sampling sequence is known as the frequency interpretation of probability. The following things should be noted:

1. An example of a sequence for which the limit defining $P(x)$ does *not* exist is the following. Suppose Ω consists of the two elements A and B. Then the sampling sequence

$$\{x_n\} = A, B, A, B, A, A, B, B, A, A, A, A, \ldots \tag{1.7}$$

   does not define a function $P(x)$. Therefore the existence of the limit defining $P(x)$ is not a completely trivial restriction on acceptable sampling sequences.
2. We have not imposed any requirement of randomness or unpredictability on the sampling sequence. Although many of the cases we will consider lead to unpredictable sampling sequences, others do not. For example, a very important case in statistical mechanics is one in which the sampling sequence is the sequence of dynamical states of an isolated mechanical system at a sequence of time instants, t_1, t_2, t_3, ... For a deterministic dynamical system these are not at all unpredictable. They will not have to be.

*This is a deviation from standard probability theory terminology where Ω is called the *sample space*. It will be far more natural for our major application which is to statistical mechanics.

All of the results of probability theory that will be used are mathematical consequences of the assumed properties of $P(x)$ and no property of randomness will be necessary in deriving those consequences.

The function $P(x)$ will be called the *fundamental microstate probability*. In most cases $P(x)$ is determined either by actually carrying out measurements on a real sampling sequence or from some sort of plausability argument, as was used previously for well-made dice. Formal probability theory then shows us how to calculate the probability of statements that depend upon x (for example, $R+B=6$) in terms of the fundamental microstate probability.

Consider now some statement or equation $q(x)$ whose truth is a function of x. The probability of $q(x)$ would naturally be defined as

$$\mathbf{P}[q] = \lim_{N\to\infty}\Big(\frac{\text{the number of times } q(x_n) \text{ is true for } n=1,\ldots,N}{N}\Big) \tag{1.8}$$

Any such statement defines a subset $Q\subset\Omega$ by $x\in Q$ if and only if $q(x)$ is true. It is then easy to show that

$$\mathbf{P}[q] = \sum_{x\in Q} P(x) \tag{1.9}$$

The relation between statements and subsets of Ω suggests that the definition of $P(x)$ be generalized to a probability $\mathbf{P}[Q]$ assigned to any subset $Q\subset\Omega$ by

$$\mathbf{P}[Q] = \lim_{N\to\infty}\Big(\frac{\text{the number of } x_n\in Q \text{ for } n=1,\ldots,N}{N}\Big) \tag{1.10}$$

A probability function that is defined for the subsets of Ω is called a *probability measure*.

Now let us see why we restricted ourselves to a system with a finite number of microstates. For finite Ω the definition of $\mathbf{P}[Q]$ obviously implies that

$$\mathbf{P}[Q] = \sum_{x\in Q} P(x) \tag{1.11}$$

and, in particular, that

$$\sum_{x\in\Omega} P(x) = \mathbf{P}[\Omega] = 1 \tag{1.12}$$

But suppose Ω consists of all the positive integers and we consider the sampling sequence $x_1, x_2, \ldots = 1, 2, \ldots$. That is, $x_n = n$. For this sampling sequence the limit defining $P(x)$ converges to zero for each $x\in\Omega$. Thus $P(x)$ is well defined. Also, using Eq. (1.10), we get $\mathbf{P}[\Omega]=1$. But then

$$\sum_{x\in\Omega} P(x) = 0+0+\cdots = 0 \neq \mathbf{P}[\Omega] = 1 \tag{1.13}$$

We see that certain properties of the probability function that can be proven always to hold for finite Ω are not automatic for infinite Ω. They must therefore be incorporated into the definition of an acceptable sampling sequence. Before describing the properties that will be needed we will introduce some elementary ideas of set theory. It is assumed that all the sets discussed here are subsets of some microstate space, Ω.

1.3 THE BASIC NOTIONS OF SET THEORY

The *union* of two sets, A_1 and A_2, written $A_1\cup A_2$ or A_1+A_2, is a set containing any element of Ω that is in at least one of the two sets. For example,

$$\{1,2,3\}\cup\{2,3,4\} = \{1,2,3\}+\{2,3,4\} = \{1,2,3,4\} \tag{1.14}$$

The *intersection* of two sets, A_1 and A_2, written $A_1\cap A_2$ or $A_1\cdot A_2$, is a set containing any element of Ω that is in both A_1 and A_2. Thus

$$\{1,2,3\}\cap\{2,3,4\} = \{1,2,3\}\cdot\{2,3,4\} = \{2,3\} \tag{1.15}$$

The notions of union and intersection generalize in an obvious way to arbitrary collections of sets. For example, given a sequence of sets, A_1, $A_2, \cdots$, then

$$x \in A_1 \cup A_2 \cup \cdots \text{ if and only if } x \in A_k \text{ for some } k \tag{1.16}$$

and

$$x \in A_1 \cap A_2 \cap \cdots \text{ if and only if } x \in A_k \text{ for every } k \tag{1.17}$$

The statement that every element of A_1 is also an element of A_2 (i.e., that A_1 is a *subset* of A_2) is written $A_1 \subset A_2$ (or else $A_2 \supset A_1$). The *complement* of a set A is written $\bar{A}$ and is a set containing all elements of Ω that are not elements of A. The *difference* between two sets, written $A_1 - A_2$, is a set containing those elements of A_1 that are not elements of A_2. It is easy to see that

$$A_1 - A_2 = A_1 \cdot \bar{A}_2 \tag{1.18}$$

and that

$$\bar{A} = \Omega - A \tag{1.19}$$

The *empty set* is always denoted $\emptyset$ and is the set containing no elements

$$\emptyset = \{\} \tag{1.20}$$

As part of the definition of the word *subset*, $\emptyset$ is considered to be a subset of every set. Two sets are said to be *disjoint* if they contain no common elements. The sets in a sequence, $A_1, A_2, \ldots$, are called *mutually disjoint* if $A_i \cap A_j = \emptyset$ whenever $i \neq j$.

1.4 THE MEASURE THEORY INTERPRETATION

We now return to the question of what are the requirements of an acceptable probability measure. A probability function $\mathbf{P}[A]$ defined for all $A \subset \Omega$, will be deemed acceptable if and only if it satisfies the following conditions

1. $\mathbf{P}[A] \geq 0$ for all $A \subset \Omega$.
2. $\mathbf{P}[\Omega] = 1$.
3. For any (finite or infinite) sequence, $A_1, A_2, \cdots$, of mutually disjoint sets

$$\mathbf{P}[A_1 \cup A_2 \cup \ldots] = \mathbf{P}[A_1] + \mathbf{P}[A_2] + \ldots \tag{1.21}$$

From (3) it is easy to show that $\mathbf{P}[\emptyset] = 0$. (Let $A_1 = \emptyset$.) Properties (1), (2), and (3) can be proven to hold whenever Ω has a finite number of elements, but if Ω contains an infinity of elements, then, as we have seen, these properties, in particular 3, are not an automatic consequence of the convergence of the limit defining $\mathbf{P}[A]$. They will be postulated to hold for any probability distribution we consider henceforth.

It is possible to dispense with the notion of a sampling sequence entirely and to base the study of probability theory completely on Postulates 1, 2, and 3. This defines a measure theory interpretation of probability. One simply postulates a function, $\mathbf{P}[A]$, defined for all subsets of a microstate space, Ω, and satisfying conditions 1, 2, and 3.* For a strictly mathematical study of probability theory the measure theory interpretation is definitely the most convenient one. However, in applying probability theory to the analysis of physical systems, the idea of a sampling sequence supplies an interpretation that is more concrete and pictorial and therefore helps in developing an intuitive understanding of the subject.

1.5 SOME EXAMPLES

Before going any further in the general theory, let us consider some simple problems in probability theory. We begin with a few of what are called *combinatorial problems*.

*If Ω contains a continuous infinity of elements, such as all the points in a two-dimensional rectangle, then there are certain technical restrictions on the collection of sets for which $\mathbf{P}[A]$ is assumed to be defined, but it would take us too deeply into measure theory to discuss them.

Question How many distinct sequences of n integers, $(I_1, I_2, \ldots, I_n)$, can be constructed if the integers are restricted to the range $1 \leq I_j \leq N$?
Answer A sequence of n objects is called an n-tuple. Two n-tuples, $(x_1, \ldots, x_n)$, and $(y_1, y_2, \ldots, y_n)$, are considered identical only if $x_j = y_j$ for all j. Otherwise they are considered to be distinct. This should be contrasted to sets of n objects. Permutation of the objects in a set does not give a distinct set. That is, the set $\{1, 3, 5\}$ is identical to the set $\{5, 1, 3\}$, but the 3-tuple $(1, 3, 5)$ is not equal to the 3-tuple $(5, 1, 3)$. Now let us answer the question. Clearly I_1 can have N possible values. For any choice of I_1, I_2 can have N possible values. Thus there are N^2 possible choices of I_1 and I_2. Proceeding in this way, it is easy to see that there are N^n possible choices or $I_1, I_2, \ldots, I_n$.

Question How many distinct sequences of length n can we construct from the first N integers if we demand that no two integers in the sequence be equal? We assume that $N \geq n$.
Answer Again, there are N choices for the first integer, I_1. Having chosen I_1, there are now only $N - 1$ choices for the second integer, I_2. Thus there are $N(N-1)$ choices of I_1 and I_2. Having chosen those two, there are only $N - 2$ choices for I_3 and so forth. Thus the number of distinct n-tuples containing n different integers drawn from the first N integers is

$$N(N-1)\ldots(N-n+1) = \frac{N!}{(N-n)!} \tag{1.22}$$

For $n = N$, this formula says that there are $N!$ permutations of N distinct objects.

Question How many distinct sets of n integers can be constructed from the first N integers?
Answer We can answer this by grouping the distinct n-tuples of the last question into classes, such that any two n-tuples in a class contain the same set of integers and are therefore permutations of one another. Clearly, each class contains $n!$ n-tuples, which is the number of permutations there are of n distinct objects. Each class corresponds to one of the sets of this question. Thus the answer to this question is given by dividing the answer to the last question by $n!$. The number obtained is called the *binomial coefficient*.

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{1.23}$$

Question How many ways can N numbered balls be grouped into three sets such that the first set contains n_1 balls, the second set n_2 balls, and the third set $n_3 = N - n_1 - n_2$ balls?
Answer There are $\binom{N}{n_1}$ ways of choosing the first set. Having done that we would be left with $N - n_1$ balls from which we can draw the second set in $\binom{N-n_1}{n_2}$ ways. The third set is then determined. Thus the answer is

$$\binom{N}{n_1}\binom{N-n_1}{n_2} = \frac{N!}{n_1!n_2!n_3!} \tag{1.24}$$

The generalization of this formula is obvious.

Question What is the probability of getting n heads in a simultaneous toss of N fair coins?
Answer A single toss is described by giving the state (H = heads or T = tails) of each of the N coins. Thus $x = (c_1, c_2, \ldots, c_N)$, where c_j is the state of coin j and $c_j = H$ or T. The total number of possible microstates is 2^N. Consider any particular microstate, $(c_1, c_2, \ldots, c_N)$. If we use the fact that coin 1 is a fair coin and we make the reasonable assumption that coin 1 is independent of the other $N - 1$ coins, then the probability of this microstate is the same as the probability of the microstate $(c_1^*, c_2, \ldots, c_N)$, where c_1^* is the "opposite" of c_1. Applying the same argument to different coins, we arrive at the conclusion that $P(x)$ has the same value for each of the 2^N possible microstates in Ω. Normalization then requires that $P(x) = 1/2^N$ for all $x \in \Omega$. Any n-head microstate defines a partition of the N coins into two sets, the n heads and the $N - n$ tails. Thus the number of n-head microstates is the same as the number of n-element sets that can be constructed from N distinct elements. Each of these microstates has a probability $1/2^N$. Therefore

$$\mathbf{P}[n \text{ heads}] = \frac{1}{2^N}\binom{N}{n} \tag{1.25}$$

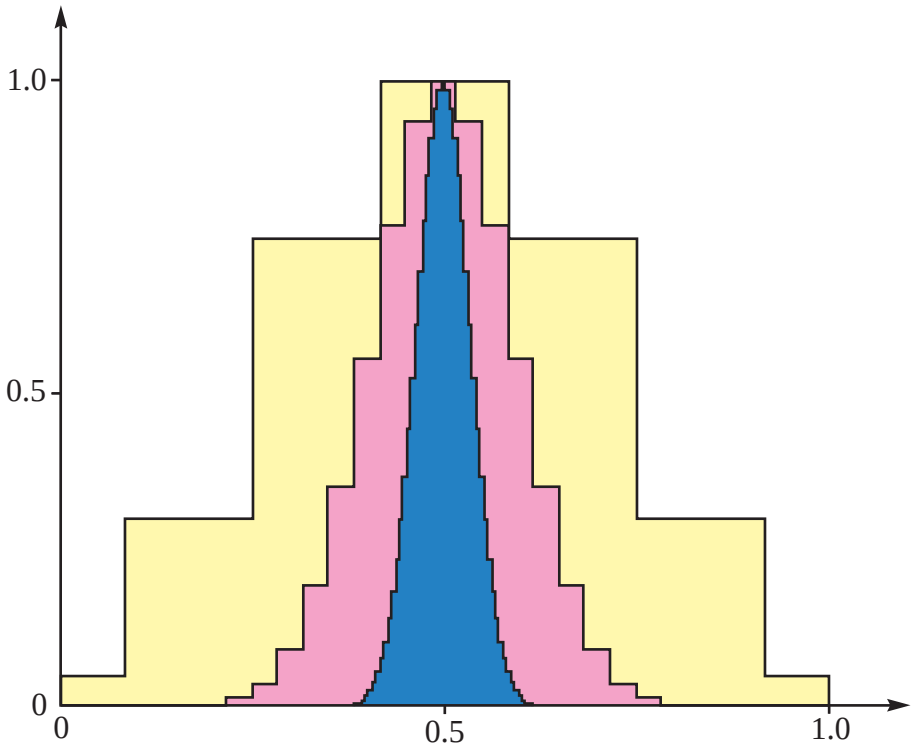

Fig. 1.1 The function $R(x)$ defined in Eq. (1.26), for the cases $N = 6$, 50, and 200. Since x must be an integer divided by N, the function is discontinuous. The wide curve is the case $N = 6$, the narrow one is the case $N = 200$. As N increases, the probability of getting significant deviations from 1/2 heads goes to zero.

For even N, the most probable number of heads is $N/2$. The ratio

$$R(x) = \frac{\mathbf{P}[xN \text{ heads}]}{\mathbf{P}[\frac{1}{2}N \text{ heads}]} \tag{1.26}$$

is plotted as a function of x for the three cases $N = 6$, 50, and 200 in Fig. 1.1. As the total number of coins increases the probability of obtaining a fraction of heads significantly different from $\frac{1}{2}$ goes to zero. This phenomenon of certain probability distributions becoming more and more sharply peaked as the number of elements in a system increases will be of central importance in our application of probability theory to systems with large numbers of particles. It is only this statistical effect that makes the macroscopic properties of large systems deterministic and predictable.

Question N weakly interacting particles move within a volume V. What is the probability of finding exactly n particles within some particular subvolume v within V?

Answer The probability that any particular particle will be found within v is $p = v/V$. Because the particle must be either inside of v or outside of v, the probability that it will be found outside of v is $1 - p$. If we ignore the interactions and therefore assume that the particles are statistically independent, then the probability that some particular set of n particles is within v while the remaining $N - n$ particles are outside v is equal to $p^n(1-p)^{N-n}$. To get the probability of finding any set of n particles within v we must multiply this by the number of ways of choosing a set of n particles from a larger set of N particles. This gives

$$\mathbf{P}[n \text{ particles in } v] = P_n = \binom{N}{n} p^n (1-p)^{N-n} \tag{1.27}$$

In the case that $v \ll V$ and $n \ll N$, this formula can be simplified by making the following approximations. (Note that $p = v/V \ll 1$.)

$$\frac{N!}{(N-n)!} = N(N-1)\cdots(N-n+1) \approx N^n \tag{1.28}$$

$$(1-p)^{N-n} \approx (1-p)^N = [(1-p)^{1/p}]^{pN} \approx e^{-pN} \tag{1.29}$$

With the definition, $\bar{n} \equiv Np$, P_n can be written as

$$P_n = \frac{\bar{n}^n}{n!} e^{-\bar{n}} \tag{1.30}$$

This is called the *Poisson distribution*. It is easy to confirm that $\bar{n}$ is actually the average value of n. That is, that

$$\bar{n} = \sum_{n=1}^{\infty} n P_n \tag{1.31}$$

If $v = V/2$, then, for obvious reasons, Eq. (1.27) for P_n becomes identical with Eq. (1.25) for the probability of obtaining n heads in a throw of N coins.

Question In a throw of three dice, what is the probability that at least one of the dice will show the number six?
Answer The probability that a given die will not show six is $\frac{5}{6}$. The dice are independent, and therefore the probability that the three will simultaneously not show the number six is $\left(\frac{5}{6}\right)^3$. The alternative is for at least one of the dice to show the number six. Therefore,

$$\mathbf{P}[\text{at least one six}] = 1 - \left(\frac{5}{6}\right)^3 = 0.42 \tag{1.32}$$

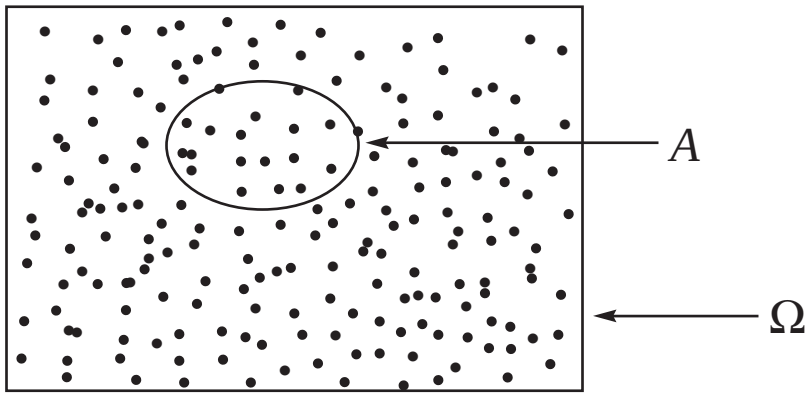

Fig. 1.2 An ensemble is a very large collection of points from the sampling sequence. The probability that the microstate will be found in the region A is the ratio of the number of ensemble points in A to the total number of points in the ensemble.

1.6 ENSEMBLES

Starting with a sampling sequence, one can construct yet another useful interpretation of the formulas of probability theory. Let us consider the case in which Ω consists of all the points in some rectangle in the x-y plane. An associated sampling sequence would be a sequence of points, (x_1, y_1), (x_2, y_2), ..., all lying within the rectangle. If, for some very large value of N, we plot the first N points in the sampling sequence, then, to a high degree of accuracy, for any region $A \subset \Omega$,

$$\mathbf{P}[A] = \frac{\text{the number of points in } A}{\text{the number of points in } \Omega} \tag{1.33}$$

The collection of a large but finite number of points, taken from the sampling sequence, is called a *statistical ensemble* (see Fig. 1.2). Of course, Eq. (1.33) for $\mathbf{P}[A]$ only becomes exact in the limit that the number of points in the ensemble goes to infinity but, in that limit one would lose the simple picture of a discrete

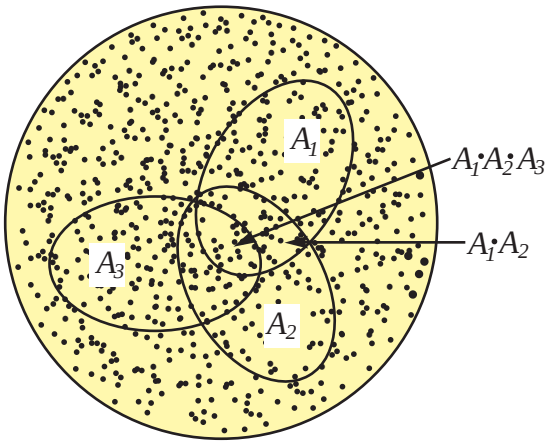

Fig. 1.3 The region $A_1 \cdot A_2$ is the overlap of the regions A_1 and A_2. It is clear that $\mathbf{P}[A_1] + \mathbf{P}[A_2]$ counts the ensemble points in $A_1 \cdot A_2$ twice, and therefore, $\mathbf{P}[A_1 + A_2] = \mathbf{P}[A_1] + \mathbf{P}[A_2] - \mathbf{P}[A_1 \cdot A_2]$. The other relations in Eq. (1.36) can be obtained in a similar way.

collection of points. Whenever the ensemble interpretation of probability is used, we will simply assume that N, the number of points in the ensemble, is sufficiently large to make any error due to finite N negligible.

1.7 PROBABILITY DENSITY

Using the ensemble picture, we can define a nonnegative function $P(x, y)$ called the *probability density*, as follows.

$$P(x, y)\, dx\, dy = \frac{\text{the number of points in } dx\, dy}{N} \tag{1.34}$$

In terms of the probability density, $\mathbf{P}[A]$ can be expressed as

$$\mathbf{P}[A] = \int_A P(x, y)\, dx\, dy \tag{1.35}$$

The phrase, *probability distribution*, will be used to refer to both a probability density function for a continuous variable, such as $P(x, y)$, or the set of discrete probabilities, $P(x)$, for a variable, x, that has only discrete values.

It will be left as an exercise for the reader (Problem 1.9), to derive, from the defining properties of a probability measure, the following facts, which are quite obvious from an ensemble picture (see Fig. 1.3).

$$\begin{aligned}
&1.\quad \mathbf{P}[A_1 + A_2] = \mathbf{P}[A_1] + \mathbf{P}[A_2] - \mathbf{P}[A_1 \cdot A_2] \\
&2.\quad \mathbf{P}[A_1 - A_2] = \mathbf{P}[A_1] - \mathbf{P}[A_1 \cdot A_2] \\
&3.\quad \mathbf{P}[A_1 + A_2 + A_3] = \sum_i \mathbf{P}[A_i] - \sum_{i<j} \mathbf{P}[A_i \cdot A_j] + \mathbf{P}[A_1 \cdot A_2 \cdot A_3]
\end{aligned} \tag{1.36}$$

1.8 PROBABILITIES OF COMPOUND STATEMENTS

A compound statement is a statement composed of one or more component statements and the logical operators *and*, *or*, or *not*. Formulas for the probabilities of compound statements follow rather directly from the meaning of the logical operators. Given a system with a microstate space Ω and given two statements $q_1(x)$ and $q_2(x)$, then the probability that the statements are both true is given by

$$\mathbf{P}[q_1 \text{ and } q_2] = \mathbf{P}[Q_1 \cdot Q_2] = \mathbf{P}[Q_1 \cap Q_2] \tag{1.37}$$

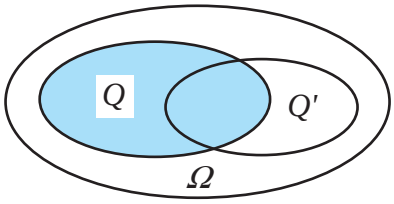

Fig. 1.4 The fraction of the ensemble points in Q that are also in Q' is the ratio of the points in the overlap area to the points in the area Q. But that is just $\mathbf{P}[Q \cdot Q']/\mathbf{P}[Q]$.

This is obvious, since q_1 is true if and only if $x \in Q_1$ and q_2 is true if and only if $x \in Q_2$. Thus "q_1 and q_2" is true if and only if $x \in Q_1 \cap Q_2$. The statement "q_1 or q_2" is true if either (or both) of the component statements is true. Therefore

$$\mathbf{P}[q_1 \text{ or } q_2] = \mathbf{P}[Q_1 + Q_2] = \mathbf{P}[Q_1 \cup Q_2] \tag{1.38}$$

It is also easy to see that

$$\mathbf{P}[\text{not } q] = \mathbf{P}[\bar{Q}] = 1 - \mathbf{P}[Q] \tag{1.39}$$

1.9 CONDITIONAL PROBABILITIES

"Given that the first two cards we are dealt in a game of poker are clubs, what is the probability that our hand will be a flush (all clubs)?" Questions of this type, involving *conditional probabilities*, are common. Their general form is: "Given statement q, what is the probability of statement q'?" Using the frequency interpretation of probability, the meaning of the question is: "If we only count those members of the sampling sequence that satisfy the relation '$q(x_n)$ is true,' what fraction of them will have $q'(x_n)$ true?" We will write the probability of q', conditioned on q as $\mathbf{P}[q'|q]$. In terms of the sampling sequence

$$\mathbf{P}[q'|q] = \lim_{N\to\infty}\left[\frac{\text{the number of times } q \text{ and } q' \text{ is true}}{\text{the number of times } q \text{ is true}}\right] \tag{1.40}$$

From this it is easy to see that

$$\mathbf{P}[q'|q] = \frac{\mathbf{P}[q' \text{ and } q]}{\mathbf{P}[q]} = \frac{\mathbf{P}[Q \cdot Q']}{\mathbf{P}[Q]} \tag{1.41}$$

This formula also follows easily from an ensemble picture. We are asking: "What fraction of the ensemble points in Q are also in Q'?" (See Fig. 1.4.)

The statement q' is said to be *statistically independent of* or *uncorrelated with* the statement q if the probability of q' being true, given that q is true, is equal to the probability of q' being true without that condition, that is, if

$$\mathbf{P}[q'|q] = \mathbf{P}[q'] \tag{1.42}$$

By using Eq. (1.41) we can see that statistical independence is a reciprocal relationship. That is, the statement "q' is statistically independent of q" implies that q is statistically independent of q'. Two statements are statistically independent if and only if

$$\mathbf{P}[q \text{ and } q'] = \mathbf{P}[q]\mathbf{P}[q'] \tag{1.43}$$

or equivalently

$$\mathbf{P}[Q \cdot Q'] = \mathbf{P}[Q]\mathbf{P}[Q'] \tag{1.44}$$

The probability of two statistically independent statements being simultaneously true is equal to the product of their separate probabilities. This is a fact that we have previously used in the examples.

Question The shots of a certain marksman on a target are found to be distributed with a probability density $P(x, y) = Ce^{-r^2}$, where $r = \sqrt{x^2 + y^2}$ is the distance from the center of the bullseye, measured in centimeters. Given that a shot is at least 1 cm high, what is the probability that it is also at least 1 cm

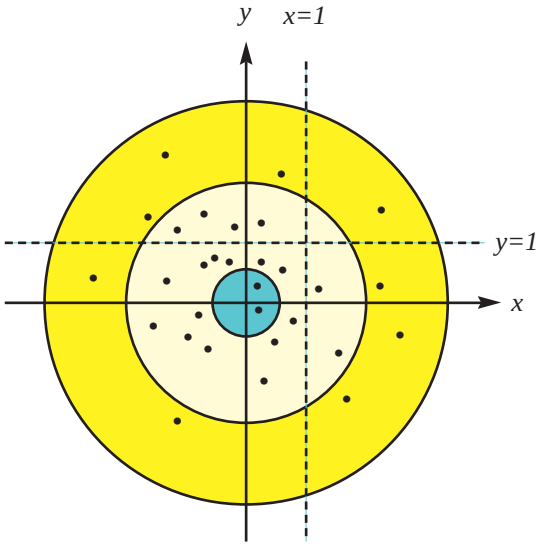

Fig. 1.5 We want the ratio of the 'ensemble points' (that is, bullet holes) in the upper right corner to the number of points above the dashed line.

to the right? **Answer** The perforated target is an excellent representation of the statistical ensemble for the system. The question is: "Of those holes that lie above the line $y = 1$, what fraction lie to the right of the line $x = 1$?" (See Fig. 1.5.) The density of holes is of the form $D(x, y) = Ae^{-x^2-y^2}$ and therefore the probability that $x > 1$, given that $y > 1$, is

$$\begin{aligned}\mathbf{P}[x > 1|y > 1] &= \frac{\int_1^\infty dx \int_1^\infty dy\, e^{-x^2-y^2}}{\int_{-\infty}^\infty dx \int_1^\infty dy\, e^{-x^2-y^2}} \\ &= \frac{\int_1^\infty dx\, e^{-x^2}}{\int_{-\infty}^\infty dx\ |e^{-x^2}} \\ &= \frac{\sqrt{\pi}}{2}\mathrm{erfc}(1) \approx 0.28\end{aligned} \tag{1.45}$$

where $\mathrm{erfc}(x)$ is the complementary error function, defined in the Mathematical Appendix. It is easy to see that $\mathbf{P}[x > 1|y > 1] = \mathbf{P}[x > 1]$, which shows that the two conditions are statistically independent. The statistical independence of the x and y distributions is a consequence of the fact that $P(x, y) = f(x)f(y)$. It can be shown (see Exercise 1.18) that the Gaussian function is the only rotationally symmetric probability distribution that yields statistically independent Cartesian coordinate distributions.

1.10 RANDOM VARIABLES

Any variable X with an associated probability distribution $P_X(x)$ is called a *random variable*.* The x and y coordinates of the shots in the above example are random variables. If X is discrete, then $P_X(x) = \mathbf{P}[X = x]$. For continuous X, $P_X(x)\,dx = \mathbf{P}[x < X < x + dx]$. Any function $Y(X)$ of the random variable X is itself a random variable with a probability distribution $P_Y(y)$. We now want to consider the problem of calculating $P_Y(y)$, given $P_X(x)$.

If X has only discrete values, then the possible values of Y must also be discrete, and the associated probability $P_Y(y)$ is given by

$$P_Y(y) = \mathbf{P}[Y = y] = \sum_{Y(x)=y} P_X(x) \tag{1.46}$$

*Why do we put the subscript X on $P_X(x)$? In this section we will be discussing the probability distributions of two or more different variables at once. If we simply write $P(x)$ and $P(y)$, rather than $P_X(x)$ and $P_Y(y)$, then we can be forced into absurd equations, such as $P(1) \neq P(1)$, where on the left we have $P(x)$ evaluated at $x = 1$ and on the right we have $P(y)$ at $y = 1$.

where the sum is over all values of x that are mapped into the value y. If X and Y are continuous, then the problem of calculating the probability density function, $P_Y(y)$, associated with the variable $Y(X)$ is more challenging. We will first consider a simple example.

Question Suppose that the two-dimensional random variable $\mathbf{X} = (X_1, X_2)$ has a probability density $P_{\mathbf{X}}(x_1, x_2)$. What is the probability density associated with the variable $R = \sqrt{X_1^2 + X_2^2}$?
Answer For any $r > 0$, the probability that $R < r$ is

$$\mathbf{P}[R < r] = \int \theta(r - R(\mathbf{x}))P_{\mathbf{X}}(\mathbf{x})\, d^2x \tag{1.47}$$

where the step function $\theta(x)$ is defined by

$$\theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases} \tag{1.48}$$

But

$$P_R(r)\, dr = \mathbf{P}[R < r + dr] - \mathbf{P}[R < r] = \frac{d\mathbf{P}[R < r]}{dr}\, dr \tag{1.49}$$

Using the fact that $d\theta(x)/\, dx = \delta(x)$, the Dirac delta function, we obtain

$$P_R(r) = \int \delta(r - R(\mathbf{x}))P_{\mathbf{X}}(\mathbf{x})\, d^2x \tag{1.50}$$

In general, if $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ is an N-dimensional random variable with a probability density $P_{\mathbf{X}}(x_1, x_2, \ldots, x_N)$ and $Y(\mathbf{X})$ is any function of $\mathbf{X}$, then

$$P_Y(y) = \int \delta(y - Y(\mathbf{x}))P_{\mathbf{X}}(x_1, x_2, \ldots, x_N)\, d^Nx \tag{1.51}$$

1.11 AVERAGES AND UNCERTAINTIES

The *average* (or *mean* or *expectation*) *value* of a random variable x with a probability distribution $P(x)$ is written $\bar{x}$ or $\langle x \rangle$ and defined as

$$\bar{x} = \langle x \rangle = \int xP(x)\, dx \tag{1.52}$$

The *root-mean-square* (rms) *uncertainty* (or *fluctuation* or *dispersion*) in x is written Δx and defined by

$$(\Delta x)^2 = \langle (x - \bar{x})^2 \rangle = \int (x - \bar{x})^2 P(x)\, dx \tag{1.53}$$

Expanding the square and remembering that $\langle x \rangle \equiv \bar{x}$, one obtains

$$(\Delta x)^2 = \langle x^2 \rangle - 2\langle x \rangle \bar{x} + \bar{x}^2 = \langle x^2 \rangle - \langle x \rangle^2$$

1.12 CHEBYSHEV'S INEQUALITY

If a random variable x has an average value $\bar{x}$ with a very small uncertainty, then it is obvious that large deviations from $\bar{x}$ must be improbable. This relationship between the uncertainty in a random variable and the probability of deviations greater than any given magnitude is precisely expressed in an inequality due to Chebyshev. It states that, for any $a > 0$,

$$\mathbf{P}[(x - \bar{x})^2 > a^2] \leq \left(\frac{\Delta x}{a}\right)^2 \tag{1.54}$$

Before giving a proof of the inequality, let us consider some of its consequences. If we choose $a = \Delta x$, then the inequality is empty in that any probability must be less than or equal to one. However, for $a = K\Delta x$, with $K > 1$, Chebyshev's inequality shows that the probability of a deviation larger that $K\Delta x$ is less than or equal to $1/K^2$. Later we will use this inequality in proving a very fundamental theorem of statistical mechanics.

The proof of Chebyshev's inequality is straightforward. We let $y(x) \equiv (x - \bar{x})^2$. The range of the random variable y is $0 \leq y < \infty$. Clearly

$$\begin{aligned}(\Delta x)^2 = \langle y \rangle &= \int_0^\infty P_y(y) y \, dy \\ &\geq \int_{a^2}^\infty P_y(y) y \, dy \\ &\geq a^2 \int_{a^2}^\infty P_y(y) \, dy \\ &= a^2 \mathbf{P}[y > a^2]\end{aligned} \tag{1.55}$$

1.13 THE LAW OF LARGE NUMBERS

Two or more random variables are said to be *independent* if they are statistically uncorrelated. For example, if X, Y, and Z are random variables and $P(x, y, z)\, dx\, dy\, dz$ is the probability of simultaneously finding X between x and $x + dx$, finding Y between y and $y + dy$, and finding Z between z and $z + dz$, then X, Y, and Z are independent if and only if

$$P(x, y, z) = P_X(x) P_Y(y) P_Z(z) \tag{1.56}$$

where P_X, P_Y, and P_Z are the probability densities associated with each variable separately.

Let $x_1, x_2, \ldots, x_N$ be N statistically independent random variables. The law of large numbers is a theorem concerning the uncertainty in the random variable

$$x \equiv \frac{x_1 + \cdots + x_N}{N} \tag{1.57}$$

If, for example, the variables $x_1, x_2, \ldots, x_N$ are the x coordinates of N identical particles, then x is the x coordinate of their center of mass. If $x_1, x_2, \ldots, x_N$ represent the states (heads or tails) of N coins, with the provision that $x_j = 1$ if the jth coin is heads and $x_j = 0$ if it is tails, then x gives the fraction of coins that are heads. The content of the theorem is that, for large N, the uncertainty in x is much smaller than the uncertainties in the individual random variables, $x_1, x_2, \ldots, x_N$.

We assume that $x_1, x_2, \ldots, x_N$ have associated probability functions, $P_1(x_1)$, ..., $P_N(x_N)$, and uncertainties, $\Delta x_1, \ldots, \Delta x_N$. We define a *mean square uncertainty*, a, by

$$a^2 = \frac{(\Delta x_1)^2 + \cdots + (\Delta x_N)^2}{N} \tag{1.58}$$

If the variables all have equal uncertainties, then $a = \Delta x_1 = \cdots = \Delta x_N$. In general, even for large N, a is of the same magnitude as the typical individual uncertainties. Because the variables are statistically independent, the joint probability density for all N variables is the product of their individual probability densities.

$$P(x_1, x_2, \ldots, x_N) = P_1(x_1) \cdots P_N(x_N) \tag{1.59}$$

We will assume that the variables x_i are continuous. The modification of the proof for the case of discrete x_i consists simply in replacing integrals by sums. The average value and dispersion of the random variable

x are given by

$$\begin{aligned}\bar{x} &= \int dx_1 \cdots \int dx_N \left(\frac{x_1 + \cdots + x_N}{N}\right) P_1(x_1) \cdots P_N(x_N) \\ &= \frac{1}{N} \int dx_1 \cdots \int dx_N \left(\sum_{i=1}^{N} x_i\right) P_1(x_1) \cdots P_N(x_N) \\ &= \frac{1}{N} \sum_{i=1}^{N} \int P_1(x_1) dx_1 \cdots \int x_i P_i(x_i) dx_i \cdots \int P_N(x_N) dx_N \\ &= \frac{\bar{x}_1 + \cdots + \bar{x}_N}{N}\end{aligned} \tag{1.60}$$

and

$$\begin{aligned}(\Delta x)^2 &= \langle (x - \bar{x})^2 \rangle \\ &= \left\langle \left(\frac{(x_1 - \bar{x}_1) + \cdots + (x_N - \bar{x}_N)}{N}\right)^2 \right\rangle \\ &= \frac{1}{N^2} \int dx_1 \cdots \int dx_N \sum_{i,j}^{N} (x_i - \bar{x}_i)(x_j - \bar{x}_j) P_1(x_1) \cdots P_N(x_N)\end{aligned} \tag{1.61}$$

Using the facts that

$$\int P_n(x_n) dx_n = 1 \tag{1.62}$$

$$\int (x_n - \bar{x}_n) P_n(x_n) dx_n = 0 \tag{1.63}$$

and

$$\int (x_n - \bar{x}_n)^2 P_n(x_n) dx_n = (\Delta x_n)^2 \tag{1.64}$$

one finds that

$$(\Delta x)^2 = \frac{\sum_n (\Delta x_n)^2}{N^2} = \frac{a^2}{N} \tag{1.65}$$

or

$$\Delta x = \frac{a}{\sqrt{N}} \tag{1.66}$$

Thus, for instance, the average of 100 independent variables has an uncertainty that is only one tenth as large as the uncertainties of the individual variables (assuming that the individual variables all have roughly equal uncertainties). The sharpening of the probability distributions illustrated in Fig. 1.1 is an example of the law of large numbers.

1.14 THE CENTRAL LIMIT THEOREM

The central limit theorem is the most powerful and important of those theorems of probability that are concerned with sums of independent random variables. It is a major extension of the law of large numbers. The central limit theorem shows that, not only does the distribution of $X = (x_1 + \cdots + x_N)/N$ become sharp, but the detailed form of $P_X(x)$ approaches a Gaussian function for large N regardless of what are the detailed forms of the separate probability densities, $P_1(x_1), \ldots, P_N(x_N)$. We will give the statement of the central limit theorem here but reserve the proof of the theorem for the Mathematical Appendix.

We consider an infinite sequence of independent random variables $x_1, x_2, \ldots$ with probability densities $P_1(x_1), P_2(x_2), \ldots$. It is no serious restriction to assume that the average value of each of the separate variables vanishes.

$$\int P_n(x_n) x_n \, dx_n = 0 \tag{1.67}$$

If X is the average of the first N variables and a is their mean square uncertainty (that is, $a^2 = (\Delta x_1^2 + \ldots + \Delta x_N^2)/N$), then, for very large N,

$$P_X(x) = (N/2\pi a^2)^{1/2} \, e^{-Nx^2/2a^2} \tag{1.68}$$

The uncertainty in X can be easily calculated from Eq. (1.68) and is

$$\Delta X = \frac{a}{\sqrt{N}} \tag{1.69}$$

This result was derived previously. What is new is the fact that, for large N, the detailed probability distribution for X is a Gaussian distribution even though no assumptions were made regarding the details of the probability distributions for the variables $x_1, x_2, \ldots, x_N$.

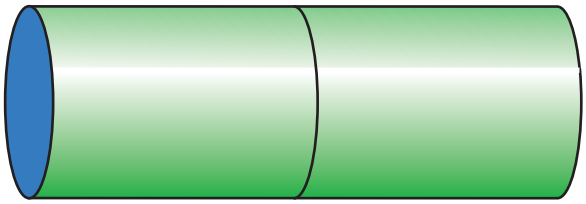

Fig. 1.6 The cylinder contains two moles of gas. The average value of the number of particles in either half is obviously one mole.

Question A cylindrical container holds two moles of an ideal gas. (Fig. 1.6) What is the probability of finding more that $(1 + 10^{-8})$ mole in the right hand half of the cylinder?

Answer We define the random variables, $\sigma_1, \sigma_2, \ldots, \sigma_N$, where $N = 2N_A$, by saying that $\sigma_n = +1$ if the nth particle is in the right-hand half of the container and $\sigma_n = -1$ otherwise. The variable σ_n takes its two possible values with equal probabilities. The average value of each σ_n is obviously zero. The square of the uncertainty in σ_n is

$$(\Delta\sigma_n)^2 = P_n(+1) \cdot (1)^2 + P_n(-1) \cdot (-1)^2 = 1 \tag{1.70}$$

The variable $X = (\sigma_1 + \cdots + \sigma_N)/N = (N_R - N_L)/N$, where N_R and N_L are the numbers of particles in the right and left sides, respectively. Equation (1.71) gives $a = 1$. The probability distribution associated with X is therefore

$$P_X(x) = (N_A/\pi)^{1/2} e^{-N_A x^2} \tag{1.71}$$

If $N_R > (1 + 10^{-8})N_A$, then $X > 10^{-8}$. Thus, letting $\epsilon = 10^{-8}$, the answer to the question is

$$p = \int_\epsilon^\infty P_X(x)\, dx = (N_A/\pi)^{1/2} \int_\epsilon^\infty e^{-N_A x^2}\, dx \tag{1.72}$$

Using the variable $u = x\sqrt{N_A}$, we can express this in terms of the complementary error function.

$$p = \frac{1}{\sqrt{\pi}} \int_{\epsilon\sqrt{N_A}}^\infty e^{-u^2}\, du = \frac{1}{2} \operatorname{erfc}(\epsilon\sqrt{N_A}) \tag{1.73}$$

Using the asymptotic approximation for the complementary error function that is valid for $z \gg 1$,

$$\operatorname{erfc}(z) \sim \frac{e^{-z^2}}{\sqrt{\pi} z} \tag{1.74}$$

we get

$$p \approx \frac{e^{-N_A \epsilon^2}}{2\epsilon\sqrt{\pi N_A}} = \frac{e^{-6 \times 10^7}}{2.7 \times 10^4} \tag{1.75}$$

This number is so small that it is physically indistinguishable from zero. Notice that the suppression of fluctuations in X is entirely a statistical effect. We have assumed that no interparticle forces exist, and therefore such things cannot be used to explain the uniformity of the density distribution.

PROBLEMS

1.1 Calculate the probability of getting the following things in one throw of dice. (a) An odd number. (b) The number 7. (c) A multiple of 3.

1.2 5 spin variables, $s_1, s_2, \ldots, s_5$, in a magnetic field are statistically independent. Each can take the values $\pm\hbar/2$ with the probabilities

$$\mathbf{P}[s = \hbar/2] = p \quad \text{and} \quad \mathbf{P}[s = -\hbar/2] = 1 - p$$

What is the probability that exactly 3 spins are up $(+\hbar/2)$?

1.3 10 computers are chosen at random from 100, of which 10 are defective. What is the probability that all 10 selected ones work?

1.4 Ignoring leap year, calculate the probability that, among 25 randomly chosen people, at least two have the same birthday.

1.5 12 books, containing a 4-volume series, are placed in random order on a shelf. What is the probability that the series is placed together and in order from left to right?

1.6 Let x be a pair of integers (m, n), each in the range from 1 to 10. The fundamental probability of x is $P(x) = (m + n)C$, where C is a constant. What is the probability that $m = n$?

1.7 How many different ways are there of seating 10 people at a round table with 12 chairs if two seatings that are related by a simple rotation are not considered as different?

1.8 Each of the 50 states supplies two U.S. senators. If N senators are chosen at random, what is the probability that they represent exactly K states? (Warning: this is a very difficult problem.)

1.9 From the assumptions listed in Section 1.4, derive Eq. (1.36).

1.10 Prove the following set identities, due to De Morgan. (a) $\overline{A \cup B} = \bar{A} \cap \bar{B}$; (b) $\overline{A \cap B} = \bar{A} \cup \bar{B}$ Feel free to use words or pictures, but do not use the set notations + or −, because there is a danger of error, due to the fact that $(A + B) - B \neq A$.

1.11 Let A and B be two subsets of Ω and let $C_1 = A$, $C_2 = B - A$, and $C_3 = \overline{A \cup B}$. Show that $\sum_i \mathbf{P}[C_i] = 1$ and $\mathbf{P}[C_i \cap C_j] = 0$ for $i \neq j$.

1.12 A (finite or infinite) sequence of statements $q_1, q_2, \ldots$, is *complete and mutually exclusive* if the probability that one and only one of the statements is true is equal to one. For example, the set of statements "today is Monday", today is Tuesday", ..., today is Sunday" is complete and mutually exclusive. Assuming that $q_1, q_2, \ldots$ are complete and mutually exclusive, prove that, for any statement q,

$$\mathbf{P}[q] = \sum_i \mathbf{P}[q|q_i]\mathbf{P}[q_i]$$

1.13 For three well-made dice (red, blue, and green), given that $R + B + G = 12$, what is probability distribution for R?

1.14 x is a random variable with a probability distribution that is uniform within the interval $(0, 1)$ and zero outside that interval. What is the probability distribution of $y = \ln(x)$?

1.15 Show that it is impossible to have a point distributed in the x–y plane with the probability density $P(x, y) = C/(x^2 + y^2 + a^2)$, where C and a are constants.

1.16 Each of the 1 Ω resistors shown in Fig. 1.7 actually has a probability density for its resistance of $P(r) = \sqrt{\alpha/\pi}\exp[-\alpha(r-1)^2]$. They are statistically independent. What is the probability distribution for the combined resistance r_{ab}? [Note: Assume that $\alpha \gg 1$ so that $P(r) \approx 0$ for the unphysical negative values of r. Then one can take all the integrals from $-\infty$ to ∞.]

Fig. 1.7

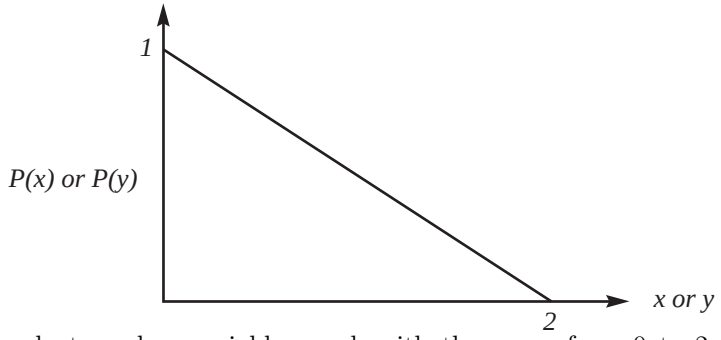

Fig. 1.8

1.17 x and y are independent random variables, each with the range from 0 to 2 and the probability distribution shown in Fig 1.8. What is the probability function associated with the variable $s = x + y$? (Note that the range of s is 0 to 4.)

1.18 Show that it is impossible for a point to be distributed over the infinite plane with a probability density of the form $P(x, y) = C/(x^2 + y^2 + 1)$.

1.19 A 1 inch needle is dropped on a sheet of paper on which parallel lines are drawn, 1 inch apart. What is the probability that it misses the lines?

1.20 A 1 inch needle is dropped on a sheet of paper on which 1 inch squares are drawn. What is the probability that it falls completely within one square?

1.21 $P(x, y)$ is uniform within the circle $x^2 + y^2 < a^2$. Calculate $P(x|y)$.

1.22 A harmonic oscillator oscillates with amplitude A. If the time t is chosen at random, what is the probability that $a \le x(t) \le a + da$? That is, find $P_x(a)$.

1.23 In a one-dimensional gas the particles have a velocity distribution $P(v) = (a/\pi)^{1/2} e^{-av^2}$. That is, the probability that a particle picked at random has velocity between v and $v + dv$ is $P(v)dv$. What is the probability that the next particle that passes the origin from left to right has velocity in the range v to $v + dv$?

1.24 A gas contains equal numbers of A particles and B particles. The A particles and B particles, respectively, have energy distributions $P_A(E)$ and $P_B(E)$, where $0 \le E < \infty$. Given that a particle has an energy larger than E_o, write an expression for the probability that it is an A particle.

1.25 Let x be a random variable that is restricted to the interval $[0, L]$. Assume that $\mathbf{P}[a < x < b] = f(b - a)$ for any $0 \le a < b \le L$. Show that $f(u) = u/L$.

1.26 Prove that, for the Poisson distribution, $\Delta n = \sqrt{\langle n \rangle}$.

The next three problems are concerned with *random walks*.

1.27 Each second a particle, which was initially at $x = 0$, jumps either left or right (each with a probability of $\frac{1}{2}$) a distance a. At time $t_n = n$ the particle is at location $x_k = ka$ with probability $P(n, k)$. Calculate $P(n, k)$ and show that, as n and k approach infinity, your result agrees with the central limit theorem.

1.28 Do Problem 1.27 for the case in which the particle jumps to the right with probability R and to the left with probability $L = 1 - R$.

1.29 Each second a particle, which was initially at $x = 0$, makes a jump. The probability distribution of the jump displacements is $p(a)$, where $-\infty < a < \infty$. Assume that $p(a) = p(-a)$ and that $\langle a^2 \rangle = \beta^2$. Use the central limit theorem to determine the probability distribution for the particle's position after a long time.

1.30 For a gas at STP (standard temperature and pressure), draw a plot of the probability that a cube of side a is completely empty as a function of a.

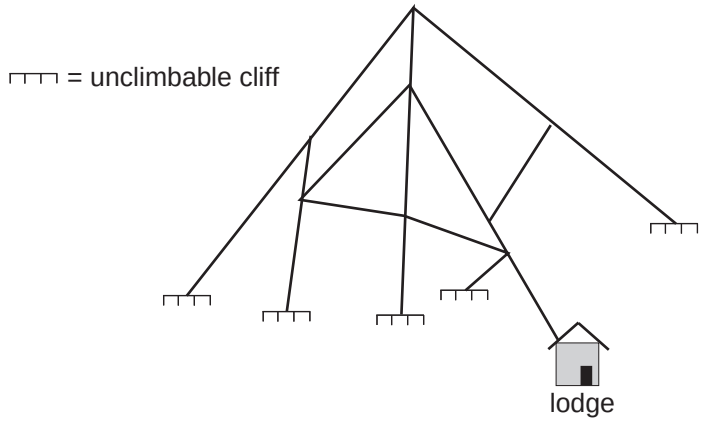

Fig. 1.9 The paths down the mountain.

1.31 Beginning at the peak and never going uphill, a hiker chooses the path randomly at each fork (see Fig. 1.9). What is the probability that she spends the night at the lodge?