

## 3

# Discrete Probability Theory

---

*The initial state will usually be a very improbable one. The system will run from this state to ever more probable states until it reaches the most probable one, which is the state of thermal equilibrium. If we apply this to the second law, we can identify the quantity that is usually called the entropy with the probability of that state.*

Ludwig Boltzmann, 1877

## 3.1 What is Probability?

The definition of probability is sufficiently problematic to have occasioned something akin to religious wars in academic departments of statistics. There are two basic factions: frequentists and Bayesians.

To understand why the definition of probability can be a problem, consider an experiment with  $N$  trials, each of which can result in a success or failure by some criterion. If  $N_s$  be the number of successes, then  $N_s/N$  is called the frequency of success for those trials.

Frequentists define the probability by the asymptotic frequency of success in the limit of an infinite number of trials,

$$p = \lim_{N \rightarrow \infty} \frac{N_s}{N}. \quad (3.1)$$

This definition looks precise and objective. Indeed, it is found as *the* definition in many physics texts. The major problem is that humans have a finite amount of time available, so that by the frequentist definition we can never determine the probability of anything. Bayesian probability provides a solution to this quandary.

The Bayesian view of probability is based on the work of Thomas Bayes (English mathematician and Presbyterian minister, 1702–1761). Bayesians define probability as a description of a person's knowledge of the outcome of an experiment, based on whatever evidence is at that person's disposal.

One great advantage of the Bayesian definition of probability is that it gives a clear meaning to a statement such as: 'The mass of a proton is  $1.672621637(83) \times 10^{-27} \text{ kg}$ ', where the '(83)' is called the uncertainty in the last digit. What does '(83)' mean? Certainly the mass of a proton is an unambiguous number that does not take on

different values for different experiments. Nevertheless, the '(83)' does have meaning as an expression of our uncertainty as to the exact value of a proton's mass.

Bayesian probability is accepted by most statisticians. However, it is in disrepute among some physicists because they regard it as subjective, in the sense that it describes what an individual knows, rather than being absolute truth. However, none of us has access to absolute truth. Bayesian statistics provide an appropriate way to describe the knowledge that an individual does have, based on the experimental results that he or she has access to.

In my opinion, the only reasonable form of objectivity that we can demand is that two observers with the same information will come to the same conclusions, and Bayesian statistics fulfills this requirement.<sup>1</sup>

The word 'probability' is used somewhat differently in statistical mechanics. I shall call it 'model probability'. It is most closely related to the frequentist meaning in that the specific values of model probabilities are assumed to be known exactly, as if we had an infinite number of trials. Of course, we really have only finite number of (imprecise) measurements, and as we shall see, we need to fix the probabilities of an enormous number of random variables (in excess of  $10^{20}$ ), so that we have no direct access to an experimental confirmation of our assumptions.

The model probability is only assumed to be valid for an equilibrium state. Indeed, all calculations will be for equilibrium, except in Chapter 22, where we discuss irreversibility. Even though we speak of thermodynamics, we limit our discussions to transitions between equilibrium states, primarily because the transitions themselves are too difficult to calculate.

The assumed form of the model probability is usually based on the plausible argument that events that are essentially alike have equal probabilities in equilibrium.

Another common argument uses the property of ergodicity, which is an additional assumption. A system is ergodic if, starting from an arbitrary point in phase space, the trajectory of the system will pass arbitrarily close to every allowable point in phase space. This trajectory is necessarily a one-dimensional subset of a  $6N$ -dimensional space. Regions in phase space that are visited equally frequently are then assigned equal probabilities. The argument to ergodicity is very popular, but it has certain weaknesses. For one thing, ergodicity by itself does not determine which regions will be visited equally frequently. Next, ergodicity has only been proven for a very limited number of systems (not the ideal gas). Such an argument also has the difficulty that it requires an enormous amount of time for a trajectory to spread through phase space—of the same order of magnitude as the Poincaré recurrence time discussed in Chapter 22.

A somewhat better argument, although still not proven for general systems, starts with an initial probability distribution that is spread over a non-zero region in phase space, which is a much more realistic assumption. If the time development of such a distribution

<sup>1</sup> This interpretation is sometimes called 'objective Bayesian probability' to distinguish it from 'subjective Bayesian probability'. Some workers do not regard the 'prior' in the Bayesian formalism as describing the information available to an experimenter before the current experiment, but rather as being an arbitrary choice. With this interpretation, Bayesian probability is subjective—and useless for physics.

leads to a unique prediction for observable quantities, that is used to justify the model probability. It has the advantage of converging *much* faster, as shown in Chapter 22 for the classical ideal gas.<sup>2</sup>

We can then use our assumed model probability to calculate the predicted outcome of an experiment. If measurements agree with our predictions, we can say that our model is consistent with the experiment. This is not the same thing as saying that our model probabilities are correct, and we will see in Chapter 22 that many different distributions can lead to the same predictions. Nevertheless, agreement with experiment is always comforting.

Statistical mechanics is based on simple assumptions, expressed as model probabilities, that lead to a wide variety of predictions in excellent agreement with experiments. How this is done is the main subject of this book.

## 3.2 Discrete Random Variables and Probabilities

We begin by defining a set of elementary events

$$A = \{a_j | j = 1, N_A\} \quad (3.2)$$

and assigning a probability  $P(a_j)$  to each event. The combination of random events and their probabilities is called a ‘random variable’. If the number of elementary events is finite or countable, it is called a ‘discrete random variable’.

The probabilities must satisfy the conditions that

$$0 \leq P(a_j) \leq 1 \quad (3.3)$$

for all  $a_j$ . An impossible event has probability zero, and a certain event has probability 1.

Random events can be anything: heads/tails, red/black, and so on. If the random events are all numbers, the random variable is called a ‘random number’.

Elementary events are defined to be exclusive—one, and only one elementary event can occur at each trial—so that the probabilities must also satisfy the normalization condition

$$\sum_{j=1}^{N_A} P(a_j) = 1. \quad (3.4)$$

<sup>2</sup> In general, the probability distribution after an irreversible process is not exactly the same as the model probability, although it gives identical predictions for the probabilities of macroscopic variables. This supports the use of model probabilities for equilibrium calculations.

To simplify notation, we will often write this equation as

$$\sum_a P(a) = 1, \quad (3.5)$$

suppressing explicit mention of the number of elementary events.

### 3.3 Probability Theory for Multiple Random Variables

If more than one thing can happen at each trial we can describe the situation with two or more sets of random events. For example, both an event from  $A$  and an event from

$$B = \{b_k | k = 1, N_B\} \quad (3.6)$$

might occur. We can then define a joint probability  $P(a_j, b_k)$ —or more simply,  $P(a, b)$ —which must satisfy

$$0 \leq P(a, b) \leq 1 \quad (3.7)$$

and

$$\sum_a \sum_b P(a, b) = 1. \quad (3.8)$$

#### 3.3.1 Marginal and Conditional Probability

Naturally, if we have  $P(a, b)$  we can retrieve the information for either  $A$  or  $B$  alone. The marginal probability of  $A$  is defined by

$$P_A(a) = \sum_b P(a, b) \quad (3.9)$$

with a similar expression for  $P_B(b)$ . A nice feature of marginal probabilities is that they automatically satisfy the positivity and normalization criteria in eqs. (3.3) and (3.4).

The name marginal probability comes from the practice of calculating it in the margins of a table of probabilities. Table 3.1 shows an example of two random variables that each take on two values.

Conditional probability is the probability of an event  $a$ , given that event  $b$  has occurred. It is denoted by  $P(a|b)$ , and is related to the full probability of both  $A$  and  $B$  occurring by the equations:

$$P(a, b) = P(a|b)P_B(b) \quad (3.10)$$

$$= P(b|a)P_A(a). \quad (3.11)$$

**Table 3.1** Example of a table of probabilities for independent random variables: The events of A (labeled ‘3’ and ‘4’) are listed down the left column, and events of B (labeled ‘1’ and ‘2’) across the top. The values of the probabilities are given in the four center squares. The margins contain the marginal probabilities, as defined in eq. (3.9).

$A \setminus B$	1	2	$P_A(a)$
3	1/2	1/4	3/4
4	1/6	1/12	1/4
$P_B(b)$	2/3	1/3	1

If  $P_B(b) \neq 0$ , the conditional probability  $P(a|b)$  can be written as

$$P(a|b) = \frac{P(a,b)}{P_B(b)}. \quad (3.12)$$

Combining eqs. (3.10) and eqs. (3.11) gives us Bayes’ theorem

$$P(a|b) = \frac{P(b|a)P_A(a)}{P_B(b)}. \quad (3.13)$$

We will discuss some of the consequences of Bayes’ theorem in Section 5.5.

### 3.3.2 Independent Variables

A particularly interesting case occurs when the probability distribution can be written as a product:

$$P(a,b) = P_A(a)P_B(b). \quad (3.14)$$

When eq. (3.14) is true, the two random variables are said to be independent because, if  $P_B(b) \neq 0$ , the conditional probability  $P(a|b)$  is then independent of  $b$ ,

$$P(a|b) = \frac{P(a,b)}{P_B(b)} = \frac{P_A(a)P_B(b)}{P_B(b)} = P_A(a), \quad (3.15)$$

and  $P(b|a)$  is independent of  $a$ .

Table 3.1 gives an example of independent random variables, and Table 3.2 gives an example of random variables that are not independent.

**Table 3.2** An example of a table of probabilities for random variables that are not independent. The table is arranged in the same way as Table 3.1.

$C \setminus D$	1	2	$P_C(c)$
3	0.5	0.25	0.75
4	0.1	0.15	0.25
$P_D(d)$	0.6	0.4	1

### 3.3.3 Pairwise Independence and Mutual Independence

If we have more than two random variables, the set of random variables might satisfy two kinds of independence: pairwise or mutual.

Pairwise independence means that the marginal distribution of any *pair* of random variables can be written as the product of the marginal distributions of the individual random variables.

Mutual independence means that the marginal distribution of any *subset* of random variables can be written as the product of the marginal distributions of the individual random variables.

It is obvious that mutual independence implies pairwise independence. Whether the converse is true is the subject of a problem in this chapter.

## 3.4 Random Numbers and Functions of Random Variables

Given an arbitrary random variable  $A = \{a_j | j = 1, \dots, N_A\}$ , we can define a numerical function on the set of elementary events,  $F = \{F(a_j) | 1, \dots, N_A\}$ . The set of random numbers  $F$ , together with their probabilities, is then also a random number.

If we introduce the Kronecker delta function

$$\delta_{x,y} = \begin{cases} 1 & x=y \\ 0 & x \neq y \end{cases} \quad (3.16)$$

then we can write the probability distribution of  $F$  compactly,

$$P_F(f) = \sum_a \delta_{f,F(a)} P_A(a). \quad (3.17)$$

As a simple illustration, consider a random variable that takes on the three values  $-1, 0$ , and  $1$ , with probabilities  $P_A(-1) = 0.2, P_A(0) = 0.3$ , and  $P_A(1) = 0.5$ . Define a

function  $F(a) = |a|$ , so that  $F$  takes on the values 0 and 1. The probabilities  $P_F(f)$  are found from eq. (3.17),

$$\begin{aligned} P_F(0) &= \sum_{a=-1}^1 \delta_{0,F(a)} P_A(a) \\ &= \delta_{0,F(-1)} P_A(-1) + \delta_{0,F(0)} P_A(0) + \delta_{0,F(1)} P_A(1) \\ &= 0 + P(0) + 0 = 0.3 \end{aligned} \quad (3.18)$$

$$\begin{aligned} P_F(1) &= \sum_{a=-1}^1 \delta_{1,F(a)} P_A(a) \\ &= \delta_{1,F(-1)} P_A(-1) + \delta_{1,F(0)} P_A(0) + \delta_{1,F(1)} P_A(1) \\ &= P_A(-1) + 0 + P_A(1) = 0.2 + 0.5 = 0.7. \end{aligned} \quad (3.19)$$

We can also use the Kronecker delta to express arbitrary combinations of random numbers to form new compound random numbers. For example, if  $X$  and  $Y$  are random numbers, and  $G(x,y)$  is an arbitrary function, we can define a new random variable,  $Z$ . The probability distribution of  $Z$  is given by a sum over all combinations of the events of  $X$  and  $Y$ , with the Kronecker delta picking out the ones that correspond to particular events of  $Z$

$$P_Z(z) = \sum_x \sum_y \delta_{z,G(x,y)} P(x,y). \quad (3.20)$$

A warning is necessary because the limits on the sums in eq. (3.20) have been suppressed. The only difficult thing about actually doing the sums is in keeping track of those limits. Since being able to keep track of limits will also be important when we get to continuous distributions, we will illustrate the technique with the simple example of rolling two dice and asking for the probability distribution of their sum.

## Example: Probability of the Sum of Two Dice

We will assume that the dice are honest, which they tend to be in physics problems, if not in the real world. Let  $X = \{x|x = 1, 2, 3, 4, 5, 6\}$  be the random number representing the outcome of the first die, with  $Y$  a corresponding random number for the outcome of the second die. The sum  $S = X + Y$ . The values taken on by  $S$  range from 2 to 12. Since all elementary events are equally likely,

$$P(x,y) = P_X(x)P_Y(y) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}, \quad (3.21)$$

and eq. (3.20) then becomes

$$P(s) = \frac{1}{36} \sum_{x=1}^6 \sum_{y=1}^6 \delta_{s,x+y}. \quad (3.22)$$

Do the sum over  $y$  first. Its value depends on whether  $s = x + y$  for some value of  $y$ , or equivalently, whether  $s - x$  is in the interval  $[1, 6]$

$$\sum_{y=1}^6 \delta_{s,x+y} = \begin{cases} 1 & 1 \leq s - x \leq 6 \\ 0 & \text{otherwise} \end{cases} \quad (3.23)$$

This places two conditions on the remaining sum over  $x$ . Only those values of  $x$  for which both  $x \leq s - 1$  and  $x \geq s - 6$  contribute to the final answer. These limits are in addition to the limits of  $x \leq 6$  and  $x \geq 1$  that are already explicit in the sum on  $X$ . Since all four of these inequalities must be satisfied, we must take the more restrictive of the inequalities in each case. Which inequality is the more restrictive depends on the value of  $s$ , as indicated in Table 3.3.

For  $s \leq 7$ , the lower bound on  $x$  is 1 and the upper bound is  $s - 1$ . For  $s \geq 7$ , the lower bound on  $x$  is  $s - 6$  and the upper bound is 6. The sums can then be evaluated explicitly,

$$P(s) = \begin{cases} \sum_{x=1}^{s-1} \frac{1}{36} = \frac{s-1}{36} & s \leq 7 \\ \sum_{x=s-6}^6 \frac{1}{36} = \frac{13-s}{36} & s \geq 7. \end{cases} \quad (3.24)$$

Even though the probability distribution of both  $X$  and  $Y$  are uniform, the probability distribution of  $S$  is not. This is an important result. Although we will be generally assuming a uniform underlying probability distribution for most microscopic quantities, we will see that the distribution of macroscopic observables will be very sharply peaked.

**Table 3.3** Determining the limits for the second sum when evaluating eq. (3.22).

	lower limit	upper limit
Limits on sum:	$x \geq 1$	$x \leq 6$
From the Kronecker delta:	$x \geq s-6$	$x \leq s-1$
More restrictive if $s \leq 7$	$x \geq 1$	$x \leq s-1$
More restrictive if $s \geq 7$	$x \geq s-6$	$x \leq 6$

It must be admitted that there are easier ways of numerically computing the probability distribution of the sum of two dice, especially if a numerical result is required. Eq. (3.20) is particularly easy to evaluate in a computer program, since the Kronecker delta just corresponds to an ‘IF’ statement. However, we will have ample opportunity to apply the method just described to problems for which it is the simplest approach. One reason why the example in this section was chosen is that the correct answer is easy to recognize, which makes the method more transparent.

### 3.5 Mean, Variance, and Standard Deviation

The mean, or average, of a function  $F(A)$ , defined on the random variable  $A$ , is given by

$$\langle F \rangle \equiv \sum_a F(a) P_A(a). \quad (3.25)$$

Similarly, the  $n$ -th moment of  $F$  can be defined as

$$\langle F^n \rangle \equiv \sum_a F(a)^n P_A(a). \quad (3.26)$$

The  $n$ -th central moment is defined by subtracting the mean before taking the  $n$ -th power in the definition

$$\langle (F - \langle F \rangle)^n \rangle \equiv \sum_a (F(a) - \langle F \rangle)^n P_A(a). \quad (3.27)$$

The second central moment plays an important role in statistical analysis and is called the variance,  $\sigma^2$

$$\sigma^2 = \langle (F - \langle F \rangle)^2 \rangle = \sum_a (F(a) - \langle F \rangle)^2 P_A(a). \quad (3.28)$$

It can also be written as

$$\sigma^2 = \langle (F - \langle F \rangle)^2 \rangle = \langle F^2 \rangle - \langle F \rangle^2. \quad (3.29)$$

The square root of the variance is called the standard deviation,

$$\sigma \equiv \sqrt{\langle F^2 \rangle - \langle F \rangle^2}. \quad (3.30)$$

The standard deviation is frequently used as a measure of the width of a probability distribution.

### 3.6 Correlation Function

Suppose we have two random numbers,  $A$  and  $B$ , and their joint probability distribution,  $P(A, B)$ , along with functions  $F(A)$  and  $G(B)$  defined on the random variables.  $F$  and  $G$  are random numbers, and we can ask questions about how they are related. In particular, we can define a correlation function,  $f_{FG}$ ,

$$f_{FG} = \langle FG \rangle - \langle F \rangle \langle G \rangle. \quad (3.31)$$

If  $F$  and  $G$  are independent random numbers, we would expect the correlation function to vanish, which it does:

$$\begin{aligned} f_{FG} &= \sum_a \sum_b F(a)G(b)P(a, b) \\ &\quad - \sum_a F(a)P_A(a) \sum_b G(b)P_B(b) \\ &= \sum_a \sum_b F(a)G(b)P_A(a)P_B(b) \\ &\quad - \sum_a F(a)P_A(a) \sum_b G(b)P_B(b) \\ &= 0. \end{aligned} \quad (3.32)$$

### 3.7 Sets of Independent Random Numbers

Given a set of random numbers  $\{F_j | j = 1, \dots, N\}$ , we are often interested in a new random number formed by taking the sum

$$S = \sum_{j=1}^N F_j. \quad (3.33)$$

We can easily calculate the mean of the random number  $S$ , which is just the sum of the means of the individual random numbers

$$\langle S \rangle = \left\langle \sum_{j=1}^N F_j \right\rangle = \sum_{j=1}^N \langle F_j \rangle. \quad (3.34)$$

If the random numbers are pairwise independent, we can also calculate the variance and the standard deviation,

$$\begin{aligned}
\sigma_S^2 &\equiv \langle S^2 \rangle - \langle S \rangle^2 & (3.35) \\
&= \sum_{j=1}^N \sum_{k=1}^N \langle F_j F_k \rangle - \sum_{j=1}^N \langle F_j \rangle \sum_{k=1}^N \langle F_k \rangle \\
&= \sum_{j=1}^N \sum_{k=1, k \neq j}^N \langle F_j \rangle \langle F_k \rangle + \sum_{j=1}^N \langle F_j^2 \rangle - \sum_{j=1}^N \langle F_j \rangle \sum_{k=1}^N \langle F_k \rangle \\
&= \sum_{j=1}^N (\langle F_j^2 \rangle - \langle F_j \rangle^2) \\
&= \sum_{j=1}^N \sigma_j^2.
\end{aligned}$$

In this derivation,  $\sigma_j^2$  denotes the variance of the  $j$ -th random number. We see that the variance of the sum of a set of *pairwise independent* random numbers is just the sum of the variances.

If the random numbers  $F_j$  all have the same mean and variance, these equations simplify further. If  $\langle F_j \rangle = \langle F \rangle$  for all  $j$ , then

$$\langle S \rangle = \sum_{j=1}^N \langle F_j \rangle = N \langle F \rangle. \quad (3.36)$$

Similarly, if  $\sigma_j^2 = \sigma^2$  for all  $j$ , then

$$\sigma_S^2 = \sum_{j=1}^N \sigma_j^2 = N\sigma^2. \quad (3.37)$$

Note that the standard deviation of  $S$  grows as the square root of the number of variables

$$\sigma_S = \sigma \sqrt{N}. \quad (3.38)$$

On the other hand, the relative standard deviation decreases with the square root of the number of variables

$$\frac{\sigma_S}{\langle S \rangle} = \frac{\sigma \sqrt{N}}{N \langle F \rangle} = \frac{\sigma}{\langle F \rangle \sqrt{N}}. \quad (3.39)$$

It might be argued that this is the most important result of probability theory for statistical mechanics. For many applications in statistical mechanics, the appropriate value of  $N$  is  $10^{20}$  or higher, so that the relative uncertainties for macroscopic quantities are generally of the order of  $10^{-10}$  or smaller. This is far smaller than most experimental errors, leading to predictions with no measurable uncertainty.

### 3.8 Binomial Distribution

A particularly important case is that of  $N$  independent, identically distributed random numbers,  $\{F_j\}$ , that can each take on the value 1 with probability  $p$  and 0 with probability  $1 - p$ . An example would be  $N$  flips of a coin, where the coin might be fair ( $p = 0.5$ ) or biased ( $p \neq 0.5$ ).

The mean and variance of each random number are easily seen to be  $\langle F \rangle = p$  and  $\sigma^2 = p(1 - p)$ . The mean and variance of the sum

$$S = \sum_{j=1}^N F_j \quad (3.40)$$

are then

$$\langle S \rangle = pN \quad (3.41)$$

and

$$\sigma_S^2 = p(1 - p)N. \quad (3.42)$$

The standard deviation is

$$\sigma_S = \sqrt{p(1 - p)N} \quad (3.43)$$

and the relative standard deviation is

$$\frac{\sigma_S}{\langle S \rangle} = \frac{\sqrt{p(1 - p)N}}{pN} = \sqrt{\frac{1 - p}{pN}}. \quad (3.44)$$

#### 3.8.1 Derivation of the Binomial Distribution

We can go further and calculate the explicit probability distribution  $P(S)$  of the sum of random numbers. This result will be extremely important in the analysis of the classical ideal gas.

The probability of a specific subset of  $n$  random variables taking on the value 1, while the remaining  $N - n$  random variables that take on the value 0 is easily seen to be

$$p^n(1-p)^{N-n}. \quad (3.45)$$

To complete the calculation, we only need to determine the number of permutations of the random variables with the given numbers of ones and zeros. This is the same as the number of ways by which  $N$  distinct objects can be put into two boxes, such that  $n$  of them are in the first box and  $N - n$  are in the second.

To calculate the number of permutations, first consider the simpler problem of calculating the number of ways in which  $N$  distinct objects can be ordered. Since any of the  $N$  objects can be first,  $N - 1$  can be second, and so on, there are a total of  $N! = N(N - 1)(N - 2)\cdots 2 \cdot 1$  permutations.

For our problem of putting objects into two boxes, the order of the objects in each box does not matter. Therefore, we must divide by  $n!$  for over counting in the first box and by  $(N - n)!$  for over counting in the second box. The final number of permutations is known as the binomial coefficient and has its own standard symbol

$$\binom{N}{n} = \frac{N!}{n!(N - n)!}. \quad (3.46)$$

Multiplying by the probability given in eq. (3.45) gives us the binomial distribution

$$P(n|N) = \frac{N!}{n!(N - n)!} p^n(1-p)^{N-n} = \binom{N}{n} p^n(1-p)^{N-n}. \quad (3.47)$$

The binomial distribution acquires its name from the binomial theorem, which states that for any numbers  $p$  and  $q$ ,

$$(p + q)^N = \sum_{n=0}^N \frac{N!}{n!(N - n)!} p^n q^{N-n} = \sum_{n=0}^N \binom{N}{n} p^n q^{N-n}. \quad (3.48)$$

Setting  $q = 1 - p$  proves that the binomial distribution in eq. (3.47) is normalized.

### 3.8.2 Useful Identities for the Binomial Coefficients

Although the evaluation of the binomial coefficients appears to be straightforward,  $N!$  grows so rapidly as a function of  $N$  that the numbers are too large for a direct application of the definition. For example, a popular spreadsheet overflows at  $N = 171$ , and my hand calculator cannot even handle  $N = 70$ .

On the other hand, the binomial coefficients themselves do not grow very rapidly with  $N$ . The following identities, which follow directly from eq. (3.46), allow us to calculate

the binomial coefficients for moderately large values of  $n$  and  $N$  without numerical difficulties,

$$\binom{N}{0} = \binom{N}{N} = 1 \quad (3.49)$$

$$\binom{N-1}{n} + \binom{N-1}{n-1} = \binom{N}{n} \quad (3.50)$$

$$\binom{N}{n+1} = \frac{N-n}{n+1} \binom{N}{n}. \quad (3.51)$$

The proofs of these identities will be left as exercises.

### 3.9 Gaussian Approximation to the Binomial Distribution

For a fixed value of  $p$  and large values of  $N$ , the binomial distribution can be approximated by a Gaussian function. This is known as the central limit theorem. We will not prove it, but we will show how to determine the appropriate parameters in the Gaussian approximation.

Consider a general Gaussian function,

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-x_o)^2}{2\sigma^2}\right]. \quad (3.52)$$

The mean and the mode (the location of the maximum) coincide

$$\langle x \rangle = x_o = x_{max}. \quad (3.53)$$

The variance is given by the second central moment

$$\begin{aligned} \langle (x-x_o)^2 \rangle &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-x_o)^2 \exp\left[-\frac{(x-x_o)^2}{2\sigma^2}\right] dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} y^2 \exp\left[-\frac{y^2}{2\sigma^2}\right] dy \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{2} \sqrt{\pi} [2\sigma^2]^{3/2} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{\pi} \sqrt{2\sigma^2} \sigma^2 \\ &= \sigma^2. \end{aligned} \quad (3.54)$$

It is now easy to find a Gaussian approximation to the binomial distribution, since both the mean and the variance are known exactly from eqs. (3.41) and (3.42),

$$\langle n \rangle = pN \quad (3.55)$$

and

$$\sigma^2 = p(1-p)N. \quad (3.56)$$

A Gaussian function with this mean and variance gives a good approximation to the binomial distribution for sufficiently large  $n$  and  $N$

$$P(n|N) \approx \frac{1}{\sqrt{2\pi p(1-p)N}} \exp\left[-\frac{(n-pN)^2}{2p(1-p)N}\right]. \quad (3.57)$$

How large  $n$  and  $N$  must be for eq. (3.57) to be a good approximation will be left as a numerical exercise.

### 3.10 A Digression on Gaussian Integrals

To derive a good approximation for  $N!$  that is practical and accurate for values of  $N$  as large as  $10^{20}$ , we need to develop another mathematical tool: Gaussian integrals.

Gaussian integrals are so important in statistical mechanics that they are well worth a slight detour. Although the formulas derived in this section can be found wherever fine integrals are sold, a student of statistical mechanics should be able to evaluate Gaussian integrals without relying on outside assistance.

The first step in evaluating the integral

$$G = \int_{-\infty}^{\infty} e^{-ax^2} dx \quad (3.58)$$

will later prove to be very useful in non-Gaussian integrals later in the book because we sometimes only need the dependence of the integral on the parameter  $a$ . Make the integral in eq. (3.58) dimensionless by the transformation  $y = x\sqrt{a}$

$$G = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-y^2} dy. \quad (3.59)$$

Note that the dependence on the parameter  $a$  appears as a simple factor in front of the integral.

To evaluate the dimensionless integral, we square it and recast the product as a two-dimensional integral

$$\begin{aligned}
 G^2 &= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-x^2} dx \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-y^2} dy \\
 &= \frac{1}{a} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy \\
 &= \frac{1}{a} \int_0^{\infty} e^{-r^2} 2\pi r dr \\
 &= \frac{\pi}{a} \left[ -e^{-r^2} \right]_0^{\infty} = \frac{\pi}{a}.
 \end{aligned} \tag{3.60}$$

The value of a Gaussian integral is therefore:

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}. \tag{3.61}$$

## 3.11 Stirling's Approximation for $N!$

As mentioned previously, a difficulty in using the binomial distribution is that  $N!$  becomes enormously large when  $N$  is even moderately large. For  $N = 25$ ,  $N! \approx 1.6 \times 10^{25}$ , and we need to consider values of  $N$  of  $10^{20}$  and higher! We would also like to differentiate and integrate expressions for the probability distribution, which is inconvenient in the product representation.

The problem is solved by Stirling's approximation, which is valid for large numbers—exactly the case in which we are interested. We will discuss various levels of Stirling's approximation, beginning with the simplest.

### 3.11.1 The Simplest Version of Stirling's Approximation

Consider approximating  $\ln N!$  by an integral.

$$\ln N! = \ln \left( \prod_{n=1}^N n \right) = \sum_{n=1}^N \ln(n) \approx \int_1^N \ln(x) dx = N \ln N - N + 1. \tag{3.62}$$

This is equivalent to the approximation

$$N! \approx N^N \exp(1 - N). \tag{3.63}$$

### 3.11.2 A Better Version of Stirling's Approximation

A better approximation can be obtained from an exact integral representation of  $N!$ ,

$$N! = \int_0^{\infty} e^{-x} x^N dx. \tag{3.64}$$

The correctness of eq. (3.64) can be shown by induction. It is clearly true for  $N = 0$ , since  $0! = 1 = \int_0^\infty e^{-x} dx$ . If it is true for a value  $N$ , then using integration by parts we can then prove that it is true for  $N + 1$ ,

$$\begin{aligned}(N+1)! &= \int_0^\infty e^{-x} x^{N+1} dx \\ &= \left[ -e^{-x} x^{N+1} \right]_0^\infty - \int_0^\infty (-e^{-x} (N+1)x^N) dx \\ &= (N+1)N!. \end{aligned}\tag{3.65}$$

The integral in eq. (3.64) can be approximated by noting that the integrand is sharply peaked for large values of  $N$ . This suggests using the method of steepest descent, which involves approximating the integrand by a Gaussian function of the form

$$g(x) = A \exp \left[ -\frac{(x - x_o)^2}{2\sigma^2} \right],\tag{3.66}$$

where  $A$ ,  $x_o$ , and  $\sigma$  are constants that must be evaluated.

We can find the location of the maximum of the integrand in eq. (3.64) by setting the first derivative of its logarithm equal to zero. For the Gaussian function, this gives

$$\frac{d}{dx} \ln g(x) = \frac{d}{dx} \left[ \ln A - \frac{(x - x_o)^2}{2\sigma^2} \right] = -\frac{(x - x_o)}{\sigma^2} = 0,\tag{3.67}$$

or  $x = x_o$ .

Comparing eq. (3.67) to the first derivative of the logarithm of the integrand in eq. (3.64) equal to zero, we find the location  $x_o$  of the maximum

$$0 = \frac{d}{dx} [-x + N \ln x] = -1 + \frac{N}{x},\tag{3.68}$$

or  $x = x_o = N$ . The value of the amplitude is then just the integrand in eq. (3.64),  $e^{-x} x^N$  evaluated at  $x = x_o = N$ , or  $A = e^{-N} N^N$ .

The second derivative of the logarithm of a Gaussian function tells us the value of the variance

$$\frac{d^2}{dx^2} \ln g(x) = \frac{d}{dx} \left[ -\frac{(x - x_o)}{\sigma^2} \right] = -\frac{1}{\sigma^2}.\tag{3.69}$$

When using this method of evaluating the variance of a Gaussian approximation, the second derivative of the logarithm of the function being approximated should be evaluated at  $x_o$ .

To find the value of  $\sigma^2$ , take the second derivative of the logarithm of the integrand in eq. (3.64) and evaluate it at  $x_o$ ,

**Table 3.4** Comparison of different levels of Stirling's approximation with exact results for  $N!$  'Stirling (simple)' refers to eq. (3.63), 'Stirling (improved)' to eq. (3.72), and 'Gosper' to eq. (3.73).

$N$	$N!$	Stirling	error	Stirling	error	Stirling	error
		(simple)		(improved)		(Gosper)	
1	1	1	0%	0.922	-7.79%	0.996	-0.398%
2	2	1.47	-26%	1.919	-4.05%	1.997	-0.132%
3	6	3.65	-39%	5.836	-2.73%	5.996	-0.064%
4	24	12.75	-47%	23.506	-2.06%	23.991	-0.038%
5	120	57.24	-52%	118.019	-1.65%	119.970	-0.025%
10	3628800	1234098	-66%	3598694	-0.83%	3628559	-0.007%
20	$2.43 \times 10^{18}$	$5.87 \times 10^{17}$	-76%	$2.43 \times 10^{18}$	-0.42%	$2.43 \times 10^{18}$	-0.002%

$$-\frac{1}{\sigma^2} = \frac{d^2}{dx^2}[-x + N \ln x] = \frac{d}{dx}[-1 + \frac{N}{x}] = -\frac{N}{x^2}, \quad (3.70)$$

or

$$\sigma^2 = \frac{x_o^2}{N} = N. \quad (3.71)$$

We can now use the formula for a Gaussian integral derived in Section 3.10,

$$N! = \int_0^\infty e^{-x} x^N dx \approx \int_0^\infty e^{-N} N^N \exp\left[-\frac{(x-N)^2}{2N}\right] dx \approx e^{-N} N^N \sqrt{2\pi N}. \quad (3.72)$$

As can be seen in Table 3.4, eq. (3.72) is a considerable improvement over eq. (3.63).

The procedure used in this section to approximate a sharply peaked function by a Gaussian is extremely useful in statistical mechanics because a great many functions encountered are of this form.

### 3.11.3 Gosper's Approximation

Finally, we mention a very interesting variant of Stirling's approximation due to Gosper,<sup>3</sup>

$$N! \approx e^{-N} N^N \sqrt{\left(2N + \frac{1}{3}\right)\pi}. \quad (3.73)$$

<sup>3</sup> Cited in <http://mathworld.wolfram.com/StirlingsApproximation.html>

### 34 Discrete Probability Theory

Table 3.4 shows that Gosper's approximation is extremely good, even for very small values of  $N$ .

#### 3.11.4 Using Stirling's Approximation

In applications to statistical mechanics, we are most often interested in the logarithm of the probability distribution, and consequently in  $\ln N!$ . This has the curious consequence that the simplest form of Stirling's approximation is by far the most useful, even though Table 3.4 shows that its accuracy for  $N!$  is very poor for large  $N$ .

From Gosper's approximation, we have a very accurate equation for  $\ln N!$

$$\ln N! \approx N \ln N - N + \frac{1}{2} \ln \left[ \left( 2N + \frac{1}{3} \right) \pi \right]. \quad (3.74)$$

In statistical mechanics we are most interested in large values of  $N$ , ranging from perhaps  $10^{12}$  to  $10^{24}$ . The relative error in using the simplest version of Stirling's approximation,  $\ln N! \approx N \ln N - N$ , is of the order of  $1/N$ , which is completely negligible even if  $N$  is as 'small' as  $10^{12}$ . For  $N = 10^{24}$ ,  $\ln N! \approx N \ln N - N$  is one of the most accurate approximations in physics.

### 3.12 Binomial Distribution with Stirling's Approximation

Since the last term in eq. (3.74) is negligible for large values of  $N$ , the most common approximation is to keep only the first two terms. If we use this form of Stirling's formula, we can write the binomial distribution in eq. (3.47) in an approximate but highly accurate form,

$$P(n|N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \quad (3.75)$$

$$\begin{aligned} \ln P(n|N) \approx & N \ln N - n \ln n - (N-n) \ln(N-n) \\ & + n \ln p + (N-n) \ln(1-p). \end{aligned} \quad (3.76)$$

Note that the contributions from the second term in Stirling's approximation in eq. (3.74) cancel in eq. (3.76).

Using Stirling's approximation to the binomial distribution turns out to have a number of pleasant features. We know that the binomial distribution will be peaked and that its relative width will be small. We can use Stirling's approximation to find the location of the peak by treating  $n$  as a continuous variable and setting the derivative of the logarithm of the probability distribution with respect to  $n$  in eq. (3.47) equal to zero,

$$\frac{\partial}{\partial n} \ln P(n|N) = 0 \quad (3.77)$$

or

$$\begin{aligned}
 0 &= \frac{\partial}{\partial n} [\ln N! - \ln n! - \ln(N-n)! + n \ln p + N - n \ln(1-p)] \\
 &= \frac{\partial}{\partial n} [N \ln N - N - n \ln n - n - (N-n) \ln(N-n) - (N-n) \\
 &\quad + n \ln p + N - n \ln(1-p)] \\
 &= -\ln n + \ln(N-n) + \ln p - \ln(1-p). \tag{3.78}
 \end{aligned}$$

This equation determines the location  $n_o$  of the maximum probability from the equation:

$$\frac{n_o}{N - n_o} = \frac{p}{1-p} \tag{3.79}$$

which has the solution

$$n_o = pN. \tag{3.80}$$

This means that the location of the peak, within the simplest of Stirling's approximations, gives the *exact* value of  $\langle n \rangle = pN$ .

### 3.13 Multinomial Distribution

An important generalization of the binomial distribution is the multinomial distribution. Instead of considering  $N_T$  particles divided among just two boxes, we divide them among  $M \geq 2$  boxes, labelled  $j = 1, \dots, M$ .

The assignment of each particle is assumed to be independent of the assignment of every other particle. If the probability of a given particle being assigned to box  $j$  is  $p_j$ , the probability of a specific subset of  $N_j$  particles being assigned to box  $j$  is

$$\prod_{j=1}^M p_j^{N_j}, \tag{3.81}$$

which is the generalization of eq. (3.45). This must be multiplied by the number of ways of choosing the set of  $N_j$  particles from  $N_T$  total particles. The argument runs parallel to that in Subsection 3.8.1. There are  $N_T!$  total permutations of the  $N_T$  particles. Divide them up into  $M$  boxes of  $N_j$  particles, where  $\sum_j N_j = N_T$ . The order of the particles in each box is irrelevant, so we must divide by a factor of  $N_j!$  for each box  $j$ . The resultant multinomial probability distribution is

$$P(\{N_j\}) = \left( \frac{N_T!}{\prod_{j=1}^M N_j!} \right) \prod_{j=1}^M p_j^{N_j}. \quad (3.82)$$

For application to the ideal gas, we will assume that the probability of finding a particular particle in the volume  $V_j$  of box  $j$  is proportional to  $V_j$ . If the total volume is  $V_T = \sum_j V_j$ , we have  $p_j = V_j/V_T$ . Inserting this into eq. (3.82) it becomes

$$P(\{N_j\}) = \left( \frac{N_T!}{\prod_{j=1}^M N_j!} \right) \prod_{j=1}^M \left( \frac{V_j}{V_T} \right)^{N_j} = \frac{N_T!}{V_T^{N_T}} \prod_{j=1}^M \left( \frac{V_j^{N_j}}{N_j!} \right). \quad (3.83)$$

This equation will be very important in the next chapter. Indeed, it is very important to all of classical statistical mechanics.

## 3.14 Problems

### Sample program to simulate rolling a single die

The sample program uses Python 2.

The computer program supplied first asks you how many times you want to roll a single die. It then chooses a random integer between 1 and 6, records its value, and repeats the process until it has chosen the number of random integers you specified. It then prints out the histogram (the number of times each integer occurred), tells you how many seconds the program took, and quits.

Histograms will play a significant role in the questions in this and other chapters. They are simply a record of the number of times each possible outcome is realized during a particular experiment or simulation. Histograms are a compact way of displaying information, as opposed to a list of all results generated. An estimate of the probability of each outcome is found by simply dividing the entries in the histogram by the number of trials.

The histogram is printed out in two formats to illustrate the possibilities. The time required to run the program is calculated and it will prove very useful. Since both ‘trials’ and ‘sides’ are integers, the division ‘trials/sides’ truncates the result and produces an integer, as is common in many computer languages. If Python 3 had been used, this division would have produced a floating point number.

Two special Python commands are used in this program: `numpy.zeros(sides,int)` and `random.random()`. The command `numpy.zeros(sides,int)` is in the ‘module’ `numpy`, which is imported in the first line of the program. It produces a vector of integer zeros. The command `random.random()` is in the ‘module’ `random`, which is imported in the second line of the program.

Text to be printed is enclosed in either single or double quotes, but they must match.

Take a few minutes and play with this program. Change any command and see what happens. One of the nice things about programming is that you can’t break the computer.

When you finish coding a program, run it several times exploring various regions of parameters. It takes very little of your time, it gives you a good overview of how the parameters affect the probability distributions, and it is the aspect of programming that is the most fun.

The following sample Python code is for the program OneDie\_PARTIAL.py.

```
import numpy
import random
from time import clock
t0 = clock()

trials = 100
print "Number of trials = ", trials

sides = 6
histogram = numpy.zeros(sides,int)
print histogram

j=0
while j < trials:
    r = int(random.random() * sides )
    histogram[r] = histogram[r] + 1
    j=j+1
print histogram

s=0
while s < sides:
    print s, histogram[s], histogram[s] - trials / sides
    s=s+1

t1 = clock()
print "\nProgram time = " , '%.10.4f' % (t1 - t0) , "seconds"
```

---

### PROBLEM 3.1

#### **Rolling a die with a computer**

For this problem, either modify the Python program (OneDie\_PARTIAL.py) supplied, or write your own from scratch, using whatever computer language you prefer.

1. Write a program (or modify the one given) to calculate and print out the histogram of the number of times each side occurs, the deviation of this number from one-sixth of the number of trials, the frequency with which each side occurs, and the deviation of this from one sixth. Hand in a copy of the code you used.
2. Show a typical print-out of your program.
3. Run the program for various numbers of random integers, starting with a small number, say 10, and increasing to a substantially larger number. The only upper limit is that the program should not take more than about one second to run. (Your time is valuable.)

**THIS IS A COMPUTATIONAL PROBLEM.****ANSWERS MUST BE ACCOMPANIED BY DATA.**

Please hand in hard copies for all data to which you refer in your answers.

4. As the number of trials increases, does the magnitude (absolute value) of the differences between the number of times a given side occurs and one-sixth of the number of trials increase or decrease?  
(Hint: This is not the same question as the next one.)
5. As you increase the number of trials, does the ratio of the number of times each side occurs to the total number of trials come closer to 1/6?

**PROBLEM 3.2****Mutual independence**

We defined the concept of pairwise independence in this chapter. There is a related concept called *mutual independence*. Consider the set of random variables

$$\{A_j | j = 1, \dots, N\}.$$

They are said to be mutually independent if for any subset of  $\{A_j\}$  containing  $n$  of these random variables, the marginal distribution satisfies the condition

$$P_{i,j,\dots,n}(a_i, a_j, a_k, \dots, a_n) = P_i(a_i)P_j(a_j)P_k(a_k) \cdots P_n(a_n).$$

Obviously, mutual independence implies pairwise independence. The question is whether pairwise independence implies mutual independence.

Provide a proof or a counter example.

**PROBLEM 3.3****A die with an arbitrary number of sides**

Suppose we have a die with  $S$  sides, where  $S$  is an integer, but not necessarily equal to 6. The set of possible outcomes is then  $\{n | n = 1, \dots, S\}$  (or  $\{n | n = 0, \dots, S - 1\}$ , your choice). Assume that all sides of the die are equally probable, so that  $P(n) = 1/S$ .

Since this is partly a computational problem, be sure to support your answers with data from your computer simulations.

1. What are the theoretical values of the mean, variance, and standard deviation as functions of  $S$ ? The answers should be in closed form, rather than expressed as a sum. (Hint: It might be helpful to review the formulas for the sums of integers and squares of integers.)
2. Modify the program you used for an earlier problem to simulate rolling a die with  $S$  sides, and print out the mean, variance, and standard deviation for a number of trials. Have the program print out the theoretical predictions in each case, as well as the deviations from theory to facilitate comparisons.
3. Run your program for two very different values of  $S$ . Are the results for the mean, variance, and standard deviation consistent with your predictions?  
(Do not use such long runs that the program takes more than about a second to run. It would be a waste of your time to wait for the program to finish.)

4. Experiment with different numbers of trials. How many trials do you need to obtain a *rough* estimate for the values of the mean and standard deviation? How many trials do you need to obtain an error of less than 1%? Do you need the same number of trials to obtain 1% accuracy for the mean and standard deviation?

#### PROBLEM 3.4

##### Independence and correlation functions

We have shown that if the random variables  $A$  and  $B$  were independent and  $F(A)$  and  $G(B)$  were numerical functions defined on  $A$  and  $B$ , then

$$\langle F(A)G(B) \rangle = \langle F(A) \rangle \langle G(B) \rangle.$$

Suppose have two random *numbers*,  $X$  and  $Y$ , and we know that:

$$\langle XY \rangle = \langle X \rangle \langle Y \rangle.$$

Does that imply that  $X$  and  $Y$  are independent?

Provide a proof or counter-example.

#### PROBLEM 3.5

##### Some probability calculations (the Chevalier de Méré's problem)

1. Suppose we roll an honest die ten times. What is the probability of *not* finding a '3' on any of the rolls.
2. Calculate the following probabilities and decide which is greater.

The probability of finding at least one '6' on four rolls is greater than the probability of finding at least one 'double 6' on 24 rolls of two dice.

Historical note:

This is a famous problem, which is attributed to the French writer Antoine Gombaud (1607–1684), who called himself the Chevalier de Méré (although according to Wikipedia he was not a nobleman). He was an avid but not terribly successful gambler, and he enlisted his friend Blaise Pascal (French mathematician, 1623–1662) to calculate the correct odds to bet on the dice rolls, described previously. Pascal's solution was one of the early triumphs of probability theory.

#### PROBLEM 3.6

##### Generalized dice

1. Modify your computer program to simulate the roll of  $N$  dice. Your program should let the dice have any number of sides, but the same number of sides for each die. The number of dice and the number of sides should be read in at the start of the program.

One trial will consist of  $N$  rolls. Your program should calculate the sum of the  $N$  numbers that occur during each trial. It should also compare the results for the mean, variance, and standard deviation of that sum with the theoretical predictions.

2. Test the calculations that we have carried out for the mean, variance, and standard deviation of the sum of the numbers on the dice. In each case, obtain data for a couple of different run-lengths.  
Investigate cases (a) and (b).
  - (a) Two dice, each with ten sides.
  - (b) Ten dice, each with twenty sides.
3. Use your program to investigate the width of the distribution for various numbers of two-sided dice. Does the width of the distribution increase or decrease with increasing numbers of dice? Do your results agree with the theory?

### PROBLEM 3.7

#### Mismatched dice

We derived the probability distribution for the sum of two dice using Kronecker deltas, where each die had  $S$  sides.

For this problem, calculate the probability distribution for the sum of two dice using Kronecker deltas, when one die has four sides, and the other has six sides.

### PROBLEM 3.8

#### Computer simulations of mismatched dice

1. Write a computer program to compute the probability distribution for the sum of two dice when each die has an arbitrary number of sides.

Run your program for dice of four and six sides.

2. Modify the computer program you wrote for the previous problem to compute the probability distribution for the sum of *three* dice when each die has an arbitrary number of sides.

Run your program once for all dice having six sides, and once for any combination you think is interesting.

### PROBLEM 3.9

#### Sums of 0s and 1s

The following questions are directly relevant to the distribution of ideal-gas particles between two boxes of volume  $V_A$  and  $V_B$ . If  $V = V_A + V_B$ , then  $p = V_A/V$  and  $1-p = V_B/V$ .

1. Prove the following identity for binomial coefficients:

$$\binom{N}{n+1} = \frac{N-n}{n+1} \binom{N}{n}.$$

2. Modify a new copy of your program to simulate the addition of an arbitrary number of independent random numbers,  $\{n_j | j = 1, \dots, N\}$ , each of which can take on the value 1 with probability  $P(1) = p$  and 0 with probability  $P(0) = 1 - p$ .

Have the program print the histogram of values generated. (To save paper, print only the non-zero parts of the histograms.)

Include a calculation of the theoretical probability from the binomial distribution, using the identity you proved at the beginning of this assignment. Have your program calculate the mean, variance, and standard deviation, and compare them to the theoretical values that we have calculated.

Have the program plot the theoretical prediction for the histogram on the same graph as your histogram. Plot the deviations of the simulation data from the theoretical predictions on the second graph. (A discussion of how to make plots in VPython is included in Section 4 of the Appendix.)

3. Run your program for the following cases, using a reasonable number of trials. Comment on the agreement between the theory and your computer experiment.
  - a.  $N = 10, p = 0.5$
  - b.  $N = 30, p = 0.85$
  - c.  $N = 150, p = 0.03$

### PROBLEM 3.10

#### Sums of 0s and 1s revisited: The Gaussian approximation

1. Modify the program you used for the simulation of sums of 0s and 1s to include a calculation of a Gaussian approximation to the binomial distribution.

Keep the curve in the first graph that shows the theoretical binomial probabilities, but add a curve representing the Gaussian approximation in a contrasting color.

Modify the second graph in your program to plot the difference between the full binomial approximation and the Gaussian approximation.

2. Run simulations for various values of the probability  $p$  and various numbers of ‘dice’, with a sufficient number of trials to obtain reasonable accuracy. (Remember not to run it so long that it wastes your time.)

Comment on the accuracy (or lack thereof) of the Gaussian approximation.

3. Run your program for the following cases, using a reasonable number of trials. Comment on the agreement between the theory and your computer experiment.

1.  $N = 10, p = 0.5$
2.  $N = 30, p = 0.85$
3.  $N = 150, p = 0.03$

### PROBLEM 3.11

#### The Poisson distribution

We have derived the binomial distribution for the probability of finding  $n$  particles in a subvolume out of a total of  $N$  particles in the full system. We assumed that the probabilities for each particle were independent and equal to  $p = V_A/V$ .

When you have a very large number  $N$  of particles with a very small probability  $p$ , you can simplify this expression in the limit that  $p \rightarrow 0$  and  $N \rightarrow \infty$ , with the product fixed at  $pN = \mu$ . The answer is the Poisson distribution:

$$P_\mu(n) = \frac{1}{n!} \mu^n \exp(-\mu).$$

Derive the Poisson distribution from the binomial distribution.

(Note that Stirling's formula is an approximation, and should not be used in a derivation or a proof.)

Calculate the mean, variance, and standard deviation for the Poisson distribution as functions of  $\mu$ .

### PROBLEM 3.12

#### Numerical evaluation of the Poisson distribution

In an earlier problem you derived the Poisson distribution,

$$P_\mu(n) = \frac{1}{n!} \mu^n \exp(-\mu).$$

1. Modify (a copy of) your program to read in the values of  $\mu$  and  $N$  and calculate the value of the probability  $p$ . Include an extra column in your output that gives the theoretical probability based on the Poisson distribution.

(Note: It might be quite important to suppress rows in which the histogram is zero, otherwise, the print-out could get out of hand for large values of  $N$ .)

2. Run your program for various values of  $\mu$  and  $N$ . How large does  $N$  have to be for the agreement to be good?