2
A Bit of Probability

2.1
Random events

As explained in Chapter 1, a Monte Carlo calculation is a numerical stochastic process; that is, a sequence of random events. While we shall not discuss the philosophical question of what random events are [1, 2], we shall assume that they do exist and that it is possible and useful to develop a computer program to produce effective equivalents of natural random events.

We must distinguish between elementary and composite events. Elementary events are those that we cannot (or do not choose to) analyze into still simpler events. Normally the result (head or tail) of flipping a coin or the result (1-6) of rolling a die are thought of as elementary events. In the case of a die, however, we might interest ourselves only in whether the number was even or odd, in which case there are two outcomes. Composite events are those defined from a number of elementary events. Examples include flipping a coin twice (with four outcomes, head-head, head-tail, tail-head, tail-tail). It is sometimes useful to talk of this pair as a single "event."

As far as one knows, random events occur in nature [3]; for example, the physical outcome of the scattering of an electron by an atom cannot be predicted with certainty. It is difficult to be sure which natural random events are "elementary," although we will simplify models of physical processes by treating some events as elementary, and on that basis build up composite events. The distinction between an elementary random event and others depends on one's state of knowledge and the depth of the analysis given to the problem. Thus, one important kind of event, "compound elastic scattering" of neutrons by nuclei, is usefully analyzed for theoretical purposes into a sequence of three elementary random events. A Monte Carlo calculation might or might not make that distinction, depending on its intent. On the other hand, "simple elastic scattering" is most likely an elementary event; that is, it is not possible to distinguish more basic stages.
Given an elementary event with a countable set of discrete random outcomes, \( E_1, E_2, \ldots, E_n, \ldots \), there is associated with each possible outcome \( E_k \) a number called a probability, \( p_k \), which can lie between 0 and 1,

\[
0 \leq p_k \leq 1.
\]

If the \( k \)th outcome never occurs, \( p_k = 0 \); if it is sure to occur, \( p_k = 1 \). Conversely, if \( p_k = 0 \), we say that the event almost surely does not occur; and if \( p_k = 1 \), the event almost surely occurs. Another notation for the probability of event \( E_k \) is

\[
P\{E_k\} = p_k.
\]

Some simple properties of the probability are the following:

1. \( P\{E_i \text{ and/or } E_j\} \leq p_i + p_j \).

2. \( E_i \) and \( E_j \) are said to be mutually exclusive events if and only if the occurrence of \( E_i \) implies that \( E_j \) does not occur, \( E_i \Rightarrow \overline{E}_j \) (not \( E_j \)), and vice versa. If \( E_i \) and \( E_j \) are mutually exclusive,

\[
P\{E_i \text{ and } E_j\} = 0,
\]

\[
P\{E_i \text{ or } E_j\} = p_i + p_j.
\]

3. A whole class of events can be mutually exclusive for all \( i \) and \( j \). When the class is exhaustive, that is, all possible events have been enumerated,

\[
P\{\text{some } E_i\} = \sum_i p_i = 1.
\]

In the following we consider a compound experiment consisting of just two elementary events. For clarity, we imagine the first to have outcomes \( \{E_i\} \) with probability \( p_{1i} \) and the second to have outcomes \( \{F_j\} \) and probabilities \( p_{2j} \), respectively. Each of \( p_{1j} \) and \( p_{2j} \) obeys statements 1, 2, and 3 above. An outcome of such a composite event is a pair \((E_i, F_j)\).

4. The probability of the specific outcome \((E_i, F_j)\) is \( p_{ij} \), called the joint probability for \( E_i \) and \( F_j \).

5. \( p_{ij} = p_{1i} \cdot p_{2j} \), if and only if events \( E_i \) and \( F_j \) are independent.

6. Suppose \( E_i \) and \( F_j \) are not independent; then the joint probability can be written

\[
p_{ij} = \frac{\sum_k p_{ik} \cdot \frac{p_{ij}}{\sum_k p_{ik}}}{\sum_k p_{ik}}.
\]

\[
p_{ij} = p(i) \cdot \frac{p_{ij}}{\sum_k p_{ik}}.
\]
2.2 Random Variables

$p(i)$ defines a new number called the marginal probability for event $E_i$, that is, the probability that $E_i$ does, in fact, occur, whatever the second event may be. Therefore,

$$\sum_i p(i) = \sum_i \sum_k p_{ik} = 1 \quad \text{and} \quad p(i) = p_{1i}.$$  

The marginal distribution for the second event, $F_j$ can be similarly be written as $\sum_i p_{ij}$.

7. The second factor of Eq. 2.1 is the conditional probability

$$p(j|i) \equiv \frac{p_{ij}}{\sum_k p_{ik}} \quad \text{(2.2)}$$

and is the probability for event $F_j$ occurring, given that event $E_i$ has occurred. The probability for some $F_j$ should be 1, and indeed

$$\sum_j p(j|i) = \sum_j \frac{p_{ij}}{\sum_k p_{ik}} = \frac{\sum_j p_{ij}}{\sum_k p_{ik}} = 1 \quad \text{for every } i.$$  

All joint probabilities can be factored into a marginal distribution and a conditional probability. This scheme can be generalized to treat the joint occurrence of three or more elementary events.

2.2 Random Variables

In many cases the outcome of a random event can be mapped into a numerical value, but in some circumstances it cannot (the probability of an event is always defined, but the assigning of a number to each outcome of a class of random events may not be useful). For example, when a photon interacts with an atom the photon may be scattered or it may cause other changes to happen within the atom. There is no useful way to assign a numerical value to correspond to the alternative changes. In simulating a queue, an empty queue could be equated with 0 length, but the meaning of the empty queue is really logical, not numerical. It implies that some other course of action must be taken. In general with simulations on the computer, the outcome of a random choice is often a logical event; it may imply that a different branch of the program is to be pursued. In the following discussion, however, we shall assume that for every elementary outcome $E_i$, there is associated a real number $x_i$. A random selection, $X$, of one of the possible values $x_1, x_2, \ldots$ is called a random variable. The probability that the value $x_i$ is chosen is given by $p_i = P\{X = x_i\}$. 
The expectation of this random variable $X$, that is, the stochastic mean value, is defined as

$$E(X) \equiv \sum_i P\{X = x_i\}x_i = \sum_i p_ix_i = \mu. \quad (2.3)$$

$\mu$ is called the expected or mean value. It is common in physics to write this as $\langle X \rangle$, and we shall often use that notation.

Consider some real-valued function $g(x_i) = g_i$, where the $x_i$, correspond to a countable set of elementary events with probabilities $p_i$. If $X$ is a random variable, then $g(X)$ is also a random variable. The expectation of $g(X)$ is defined as

$$E(g(X)) = \langle g(X) \rangle = \sum_i p_i g(x_i). \quad (2.4)$$

This may be illustrated by analyzing the flipping of a coin, assigning 1 to heads, 0 to tails and using two different functions $g_1(X)$ and $g_2(X)$.

<table>
<thead>
<tr>
<th>Event</th>
<th>$p_i$</th>
<th>$x_i$</th>
<th>$g_1(x_i) = 1 + 3x_i$</th>
<th>$g_2(x_i) = \frac{1+3x_i}{1+x_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$</td>
<td>heads</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>$E_2$</td>
<td>tails</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Expected Values $\langle X \rangle = \frac{1}{2} \quad \langle g_1(X) \rangle = \frac{5}{2} \quad \langle g_2(X) \rangle = \frac{4}{2}$

From the definition of the expected value of a function, we have the property that

$$\langle \text{constant} \rangle = \text{constant}$$

and that for any constants $\lambda_1$, $\lambda_2$ and two functions $g_1$, $g_2$,

$$\langle \lambda_1 g_1(X) + \lambda_2 g_2(X) \rangle = \lambda_1 \langle g_1(X) \rangle + \lambda_2 \langle g_2(X) \rangle. \quad (2.5)$$

In the table above $g_1$ is a linear function of $X$, so that

$$\langle g_1(X) \rangle = g_1(\langle X \rangle).$$

This is not true for the nonlinear function $g_2(X)$.

An important application of expected values is to the powers of $X$. The $n$th moment of $X$ is defined as the expectation of the $n$th power of $X$,

$$\langle X^n \rangle \equiv \sum_i p_i x_i^n; \quad (2.6)$$
so, for example,
\[ \langle X^2 \rangle = \sum_i p_i x_i^2. \]

The central moments of \( X \) are given by
\[ \langle g_a(X) \rangle \equiv \langle (X - \mu)^n \rangle = \sum_i p_i (x_i - \langle X \rangle)^n. \quad (2.7) \]

The second central moment has particular significance,
\[ \langle (X - \mu)^2 \rangle = \langle (X - \langle X \rangle)^2 \rangle = \sum_i p_i (x_i - \mu)^2 = \sum_i p_i x_i^2 - \langle X \rangle^2 = \langle X^2 \rangle - \langle X \rangle^2, \quad (2.8) \]

and is called the variance of \( X \) or \( \text{var}\{x\} \). The square root of the variance is a measure of the dispersion of the random variable. It is referred to as the standard deviation and sometimes the standard error. The variance of a function of the random variable, \( g(X) \), can be determined as
\[
\text{var}\{g(X)\} = \langle (g(X) - \langle g(X) \rangle)^2 \rangle = \sum_i p_i g^2(x_i) - \langle \sum_i p_i g(x_i) \rangle^2 \quad (2.9)
\]
\[ = \langle g(X)^2 \rangle - \langle g(X) \rangle^2. \]

Consider two real-valued functions, \( g_1(X) \) and \( g_2(X) \). They are both random variables, but they are not in general independent. Two random variables are said to be independent if they derive from independent events.

As we have seen in Eq. 2.5, the expectations of a linear combination is the linear combination of the expectations. This result does not require that \( g_1(x) \) and \( g_2(x) \) be independent. The effect of statistical dependence will be seen in the variance of a linear combination of the two functions,
\[
\text{var}\{\lambda_1 g_1(X) + \lambda_2 g_2(X)\} = \lambda_1^2 \text{var}\{g_1(X)\} + \lambda_2^2 \text{var}\{g_2(X)\} \quad (2.10)
\]
\[ + 2\lambda_1 \lambda_2 \langle g_1(X) g_2(X) \rangle - \lambda_1 \lambda_2 \langle g_1(X) \rangle \langle g_2(X) \rangle. \]

Let \( X \) and \( Y \) be random variables, the expectation of the product is
\[ \langle XY \rangle = \sum_{i,j} p_{ij} x_i y_j. \quad (2.11) \]

If \( X \) and \( Y \) are independent, \( p_{ij} = p_{1i} p_{2j} \) and
\[ \langle XY \rangle = \sum_i p_{1i} x_i \sum_j p_{2j} y_j = \langle X \rangle \langle Y \rangle. \]
The expectation of the product is now the product of the expectations. Let 
\( g_1(X) = X \) and \( g_2(Y) = Y \) in Eq. 2.11, then the bracketed quantity would 
vanish and
\[
\text{var}\{\lambda_1 X + \lambda_2 Y\} = \lambda_1^2 \text{var}\{X\} + \lambda_2^2 \text{var}\{Y\}. \tag{2.12}
\]

When \( X \) and \( Y \) are not necessarily independent, we introduce a new 
quantity, the covariance, which is a measure of the degree of independence of the 
two random variables \( X \) and \( Y \):
\[
\text{cov}\{X, Y\} = \langle XY \rangle - \langle X \rangle \langle Y \rangle. \tag{2.13}
\]
The covariance equals 0 when \( X \) and \( Y \) are independent and 
\[
\text{cov}\{X, X\} = \text{var}\{X\}.
\]
Note that zero covariance does not by itself imply independence of the random 
variables. The following simple example illustrates that even functional 
dependence can still yield a zero covariance. Let \( X \) be a random variable that 
may be -1, 0, or 1 with equal probabilities, and define \( Y = X^2 \). Obviously,
\[
\langle X \rangle = 0, \\
\langle XY \rangle = \langle X^3 \rangle = 0,
\]
so \( \text{cov}\{XY\} = \langle XY \rangle - \langle X \rangle \langle Y \rangle = 0. \)

The covariance can have either a positive or negative value. Another quantity 
derived from the covariance is the correlation coefficient,
\[
\rho(X, Y) = \frac{\text{cov}\{X, Y\}}{\text{var}\{X\} \text{var}\{Y\}}^{1/2} \tag{2.14}
\]
so that
\[-1 \leq \rho(X, Y) \leq 1.\]
Since the covariance can be positive or negative, the variance of a linear 
combination of two dependent random variables can be greater or less than the 
variance if the variables were independent [cf. Eq. 2.11]. A Monte Carlo calculation 
can try to take advantage of negative correlation as a means of reducing 
the variance, as will be discussed in Chapter 4.

2.2.1
The Binomial Distribution

Consider two events \( E_0 \) and \( E_1 \) that are mutually exclusive and exhaustive:
\[
P\{E_1\} = p, \quad x_1 = 1, \\
P\{E_0\} = 1 - p, \quad x_0 = 0. \tag{2.15}
\]
Let $X_i$ be the random variable that is the $i^{th}$ outcome of a series of $N$ such events. The expectations of $X_i$ and its square become

$$
E(X_i) = p \cdot 1 + (1 - p) \cdot 0 = p,
$$

$$
E(X_i^2) = p,
$$

and the variance is then

$$
\text{var}\{X_i\} = p - p^2 = p(1 - p).
$$

Each outcome is either 0 or 1, and we set $X$ to be the sum of the $N$ outcomes $\{X_i\}$,

$$
X = \sum_{i=1}^{N} X_i.
$$

The probability that $X = n$ is the probability that $n$ of the $X_i$ were 1 and $N - n$ were 0. That is,

$$
P\{X = n\} = \binom{N}{n} p^n (1 - p)^{N-n} \quad n = 0, 1, \ldots, N. \tag{2.16}
$$

This is the binomial distribution. $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ is the binomial coefficient, which counts the number of different ways in which the $n$ $E_1$’s and the $(N - n) E_0$’s may occur. The expected value of $X$ is,

$$
\langle X \rangle = \sum_{n=0}^{N} n \binom{N}{n} p^n (1 - p)^{N-n} = Np. \tag{2.17}
$$

This may be verified by direct computation of the sum, or by noting that the expected value of $X$ is the sum of the expected value of all the $X_i$. The variance of $X$ is easily determined; since the $X_i$ are independent, the result in Eq. 2.12 may be employed,

$$
\langle (X - Np)^2 \rangle = \sum_{i=1}^{N} \text{var}\{X_i\} = \sum_{i=1}^{N} p(1 - p) = Np(1 - p). \tag{2.18}
$$

### 2.2.2

**The Geometric Distribution**

Suppose we carry out a certain experiment repeatedly and independently where there are only two outcomes: failure or success. If the outcome is a failure, the experiment will be repeated; otherwise, we stop the procedure.
Let the random variable \( X \) of interest be the number of experiments we have done until success. Let \( q \) be the probability of failure in one experiment, and \( p = 1 - q \) be the probability of success.

Then

\[
P\{X = n\} = q^{n-1}p, \; n = 1, 2, \ldots,
\]

The average number of experiments that will be carried out is

\[
\langle X \rangle = \sum_{n=1}^{\infty} nq^{n-1}p = \frac{p}{(1-q)^2} = \frac{1}{p}.
\]

This is obtained from the identity

\[
\sum_{n=1}^{\infty} nx^{n-1} = \frac{d}{dx} \left( \sum_{n=0}^{\infty} x^n \right) = \frac{d}{dx} \left( \frac{1}{1-x} \right) = \frac{1}{(1-x)^2}
\]

The above identity can also be used to prove 2.17 by taking the \( N^{th} \) derivative of the sum.

The variance of \( X \) can be calculated as

\[
\text{var}\{X\} = \langle X^2 \rangle - \langle X \rangle^2
\]

\[
= \left( \frac{2}{p^2} - \frac{1}{p} \right) - \frac{1}{p^2}
\]

\[
= \frac{1}{p^2} - \frac{1}{p}
\]

As an example, in particle transport problems, the number of collisions one particle makes follows this distribution if the medium is infinite, homogeneous, and if the relative probability of different outcomes is constant.

2.2.3
The Poisson Distribution

A random variable \( X \) is said to follow a Poisson distribution when

\[
P\{X = n\} = \frac{\lambda^n}{n!} e^{-\lambda}, \; n = 0, 1, \ldots,
\]

where \( \lambda \) is a parameter of the distribution. It is easy to find that

\[
\langle X \rangle = \lambda, \quad \text{var}\{X\} = \lambda.
\]

This distribution is fundamental in the theory of probability and stochastic processes. It is of great use in applications such as research into queuing service systems and similar discrete event problems.
2.3 Continuous Random Variables

In the previous discussions we have assumed that the random events belonged to a discrete, countable set. Probabilities can be associated with continuous variables as well, giving rise to distribution functions. Such distributions are present both in nature and in artificial stochastic processes. As an example, consider the scattering of a photon by an atom. The angle at which the photon is scattered has values that are continuous between $0^\circ$ and $180^\circ$ with some angular intervals occurring more often than others.

Given that $x$ is a real, continuous number,

$$-\infty < x < \infty,$$

a distribution function (or cumulative distribution function) may be defined as

$$F(x) \equiv P\{a \text{ random selection of } X \text{ gives a value less than } x\} = P\{X \leq x\}. \quad (2.19)$$

Suppose $x_1 < x_2$. Then $x_1 < X \leq x_2$, and $X \leq x_1$ are mutually exclusive and exhaustive for events in which $X \leq x_2$. Thus

$$P\{x_1 < X \leq x_2\} + P\{X \leq x_1\} = P\{X \leq x_2\}$$

and

$$1 \geq P\{x_1 < X \leq x_2\} = P\{X \leq x_2\} - P\{X \leq x_1\} \geq 0. \quad (2.20)$$

Therefore $F(x)$ is a nondecreasing function of its argument. We conclude that $F(-\infty) = 0$ and $F(\infty) = 1$. Furthermore, the difference $P\{X \leq x_2\} - P\{X \leq x_1\}$ may be written as $F(x_2) - F(x_1)$.

The distribution function may have intervals on which it is differentiable; in these intervals the probability distribution function (pdf) may be defined as

$$f(x) \equiv \frac{dF(x)}{dx} \geq 0. \quad (2.22)$$

If $F(x)$ is not continuous, discrete values of the distribution function are singled out at the discontinuities. For example, imagine that $F(x)$ is piecewise constant everywhere except at a countable number of places (Figure 2.1); the distribution now describes a discrete set of random variables. Formally we may use the Dirac delta function to write

$$f(x) = \sum \delta(x - x_i) \times p_i$$

$(2.23)$
where \( p_i \) is the jump of the distribution function at \( x_i \). This emphasizes the fact that \( f(x) \) need not be bounded.

Consider the distribution function shown in Figure 2.2. \( F(x) = 0 \) for all \( x \leq 0 \), so no \( X \) will be chosen there and the pdf = 0. In region II, \( 0 < x < 1 \) and \( F(x) = x \); the probability of selecting an \( X \) in the interval \( 0 \leq x_1 < X \leq x_2 \leq 1 \) is \( x_2 - x_1 \) (see (2.21)). In this range, the pdf = 1, so that selecting a particular \( X \) is as likely as any other on \((0, 1)\). For \( x > 1 \), \( F(x) = 1 \), and the probability of choosing an \( X \) in this region is 0.

A slightly different example is the discontinuous distribution function in which \( F(x) \) has a step discontinuity from 0 to \( \frac{1}{2} \) at \( x = 0 \) shown in Figure 2.3. No matter how small an interval is chosen containing \( x = 0 \), the probability of choosing an \( X \) in this interval is greater than or equal to \( \frac{1}{2} \), so there is a finite probability, \( 1/2 \), for finding \( X \) exactly 0. For any other value of \( x \), the pdf is continuous. We have a combination of a discrete and a continuous pdf. Such combinations occur in nature, as for example when an atom can undergo a radiative transition to either a continuum or a discrete level. The light spectrum, which may be considered as a pdf, will be a mixture of a continuous and a discrete part.
2.4 Expectations of Continuous Random Variables

Let $f(x)$ be the pdf of a continuous random variable $x$. It has the normalization property

$$\int_{-\infty}^{\infty} f(x) \, dx = F(\infty) = 1.$$
The mean value of $x$ is defined as

\[ E(X) = \int_{-\infty}^{\infty} x \, dF(x) = \int_{-\infty}^{\infty} x f(x) \, dx, \]  

(2.24)

The expected value of any function of the random variable is defined\(^*\) as,

\[ E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) \, dx, \]

and in particular,

\[ E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \, dx. \]  

(2.25)

From Eqs. 2.24 and 2.25, the variance of $X$ may be defined as in the discrete case

\[ \text{var}\{X\} = E((X - E(X))^2) = E(X^2) - [E(X)]^2 = \langle X^2 \rangle - \langle X \rangle^2. \]  

(2.26)

The variance of a function of $X$ becomes

\[ \text{var}\{g(X)\} = E(g^2(X)) - [E(g(X))]^2. \]  

(2.27)

The variance has the following properties:

1. For a random variable $C$, which is constant (i.e., the random variable equals $C$ with probability 1),

\[ \text{var}\{C\} = 0. \]

2. For a constant $C$ and random variable $X$,

\[ \text{var}\{CX\} = C^2 \text{var}\{X\}. \]

3. For independent random variables $X$ and $Y$,

\[ \text{var}\{X + Y\} = \text{var}\{X\} + \text{var}\{Y\}. \]

In Table 2.1, $F(x), f(x) = F'(x)$, the mean and variance are given for a few representative and interesting distributions. In particular the mean value of a

\(^*\) Since $g(X)$ is a random variable, it has its own pdf; we will show below how to compute it. Given this distribution, the expectation of $g(X)$ could be alternatively defined in the same way as in 2.24. It can be proved that both the definitions will be the same. [4]
random variable drawn from a normal distribution centered around $\mu$ is $\mu$. This result is easily proved by considering the following integral

$$\frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu) \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] dx,$$

which equals 0 since the integrand changes sign on reflection of $x$ about $\mu$. The integral may be rewritten as

$$\frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] dx = \mu \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] dx.$$

The integral on the right-hand side is just the integral of the pdf and it equals 1. The mean value of $X$ is seen to be $\mu$. The variance is $\sigma^2$.

A less well-behaved example is provided by the Cauchy or Lorentz function. The mean value of a random variable sampled from a Cauchy distribution is

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{ax}{a^2 + x^2} dx.$$

To evaluate this improper integral using elementary calculus, the infinite endpoints of integration are replaced by finite quantities $b$ and $b'$, and the behavior of the integrand as $b$ and $b'$ approach infinity is considered. The integrand clearly diverges unless $b = b'$, and then the mean value is 0. This suggests that the mean value of a series of random variables may be undefined unless the variables are chosen in some special way. The variance of a random variable sampled from a Cauchy distribution is infinity since the integral

$$\langle X^2 \rangle = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{ax^2}{a^2 + x^2} dx$$

diverges no matter how it is evaluated. In spite of the infinite variance, the Cauchy distribution can be sampled and used as needed.

Note that for each distribution there is a length scale, called variously $a$, $\lambda^{-1}$, or $\sigma$. As the scale becomes small, the normalization of the pdf grows large, inversely as the length scale, so as to ensure $\int f(x) \, dx = 1$. For those distributions with a standard deviation $\sigma$, the width is proportional to $\sigma$. In the Cauchy distribution, $a$ is a measure of the width.

2.5

Bivariate continuous random distributions

A joint probability may be defined for continuous distributions,

$$F(x, y) \equiv P\{X \leq x, Y \leq y\};$$

(2.28)
$F(x, y)$ is termed a **bivariate distribution** function. The associated bivariate probability distribution function is

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$ (2.29)

and the expected value of any function of random variables $X, Y$ is

$$E(g(X, Y)) = \langle g(X, Y) \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) g(x, y) \, dx \, dy.$$ (2.30)

The covariance and correlation coefficient, $\text{cov}\{X, Y\}$ and $\rho(X, Y)$ for continuous random variables are defined as in the discrete case, replacing sums by integrals.

If $X$ and $Y$ are correlated, it is useful to write the joint pdf as

$$f(x, y) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y) \, dy} \int_{-\infty}^{\infty} f(x, y) \, dy.$$ (2.31)

We have

$$P\{X \leq x\} = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(t, y) \, dy \, dt = \int_{-\infty}^{x} m(t) \, dt ,$$

where

$$m(x) \equiv \int_{-\infty}^{\infty} f(x, y) \, dy ,$$ (2.32)

is called the marginal probability distribution function for $x$. The first factor in Eq. 2.31 is the conditional probability; that is, given an $X$, a $Y$ may be chosen from

$$f(y|x) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y) \, dy} = \frac{f(x, y)}{m(x)} .$$ (2.33)

That is, when the marginal and conditional functions can be determined, sampling a bivariate distribution requires simply sampling two univariate distributions.

The relationship in Eq. 2.31 is easily generalized to handle more than two correlated random variables, and sampling the multivariate distribution will then involve sampling the sequence of univariate distributions so defined. What happens to the marginal pdf and the conditional probability when $X$ and $Y$ are independent (i.e., $f(x, y) = f_1(x)f_2(y)$) is left for the reader.

The expectation of the conditional probability, called the **conditional expectation** of $Y$, for fixed $X$ is

$$E(Y|X) = \int_{-\infty}^{\infty} y f(y|x) \, dy = \frac{\int_{-\infty}^{\infty} y f(x, y) \, dy}{\int_{-\infty}^{\infty} f(x, y) \, dy} = \frac{\int_{-\infty}^{\infty} y f(x, y) \, dy}{m(x)} .$$
The conditional expectation \( E(Y|X) \) is a function of the random variable \( X \) and is itself a random variable. The expectation of \( E(Y|X) \) is

\[
E(E(Y|X)) = \int_{-\infty}^{\infty} E(Y|X) \, m(x) \, dx.
\]

Upon substituting in the definition for \( E(Y|X) \)

\[
E(E(Y|X)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) \, dy \, dx
= E(Y).
\]

A more general result for a function \( g(X,Y) \) is

\[
E(E(g(X,Y)|X)) = E(g(X,Y)).
\]

This result is very useful in the discussion of the "method of expected values" discussed in Section 4.2.

2.6
Sums of random variables: Monte Carlo quadrature

Suppose that the random variables \( X_1, X_2, \ldots, X_i, \ldots \) are all drawn at random, but not necessarily independently, from the probability distribution function \( f(x) \). Let \( g_i \) be a (possibly) different function of \( X_i \) and \( \lambda_i \) a real number.

Define the function \( G \):

\[
G \equiv \sum_{i=1}^{N} \lambda_i g_i(X_i), \quad (2.34)
\]

The expectation value of \( G \) is

\[
E(G) = \langle G \rangle = E \left( \sum_{i=1}^{N} \lambda_i g_i(X_i) \right)
= \sum_{i=1}^{N} \lambda_i \langle g_i(X) \rangle \quad (2.35)
\]

since the expectation value is a linear operation. If all the \( X_i \) are independent, then the variance of \( G \),

\[
\text{var}\{G\} = \langle G^2 \rangle - \langle G \rangle^2,
\]

becomes

\[
\text{var}\{G\} = \sum_{i=1}^{N} \lambda_i^2 \text{var}\{g_i(X)\}. \quad (2.36)
\]
Let \( \lambda_i = 1/N \) and all the \( g_i(x) \) be identical and equal to \( g(x) \); then the expectation value of \( G \) becomes

\[
E(G) = E\left( \frac{1}{N} \sum_{i=1}^{N} g(X_i) \right) = \frac{1}{N} \sum_{i=1}^{N} \langle g(X) \rangle = \langle g(X) \rangle.
\] (2.37)

The function \( G \), which is the arithmetic average of the \( g(x) \), has the same mean as \( g(x) \). \( G \) is said to be an "estimator" of \( \langle g(X) \rangle \). More generally an expression \( G \) is an estimator of a quantity, such as \( \int g(x) f(x) \, dx \), if its mean \( \langle G \rangle \) is a useful approximation of that quantity. Here, "useful" is intended to cover a broad range of meanings, and is best understood in the context of specific computations. We will discuss this more fully below and in connection with some applications.

The variance of \( G \) in equation 2.36 becomes

\[
\text{var}\{G\} = \text{var}\left\{ \frac{1}{N} \sum_{i=1}^{N} g(X_i) \right\} = \sum_{i=1}^{N} \frac{1}{N^2} \text{var}\{g(X)\}
\]

\[
= \frac{1}{N} \text{var}\{g(X)\}.
\] (2.38)

That is, as \( N \), the number of samples of \( X \), increases, the variance of the mean value of \( G \) decreases as \( 1/N \). This result leads to the central idea of Monte Carlo evaluation of integrals; that is, an integral may be estimated by a sum

\[
\langle g(X) \rangle = \int_{-\infty}^{\infty} g(x) f(x) \, dx = E\left( \frac{1}{N} \sum_{i=1}^{N} g(X_i) \right).
\] (2.39)

To use the relation given in Eq.2.39: select a series of random variables, \( X_i \), from \( f(x) \); evaluate \( g(x) \) for each \( X_i \). The arithmetic mean of all the values of \( g(X_i) \) is an estimate of the integral, and the variance of this estimate decreases as the number of terms increases.

### 2.7 Distribution of the mean of a random variable: A fundamental theorem

In the discussion that follows on estimating integrals, it is assumed that the variance of the random variable always exists. If the variance does not exist, the mean value will converge, although more slowly. Alternatively, it will usually be possible to recast the sampling so that the variance does exist. A general method for doing this, along with several examples, will be given in Chapter 3.
2.7 Distribution of the mean of a random variable: A fundamental theorem

The most general result of the kind we need is the "law of large numbers" of probability theory. Suppose the random variables \( X_1, X_2, \ldots, X_N \) are independent and all drawn from the same distribution. These are called independent, identically distributed (or iid) random variables. Then the expectation of each \( X \) is \( \mu \). As \( N \to \infty \), the average value of the \( \{X_i\} \),

\[
\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i
\]

converges to \( \mu \) almost surely:

\[
P\{ \lim_{N \to \infty} \bar{X}_N = \mu \} = 1.
\]

There are stronger or weaker statements that can be made, but we shall not pursue them. [5]

The implication of the law of large numbers is that the mean of \( N \) sampled random variables converges (in probability) to its expected value. In order to estimate the speed of convergence, we need stronger assumptions. The most important way of strengthening the hypothesis is to assume that the variance exists, which we do in the following.

Assume that an estimator \( G \), its mean \( \langle G \rangle \), and variance \( \text{var}\{G\} \) all exist. Then the Chebychev inequality is

\[
P\left( |G - \langle G \rangle| \geq \frac{\text{var}\{G\}}{\delta} \right) \leq \frac{\delta}{\delta}, \quad (2.40)
\]

where \( \delta \) is any positive number. This inequality could be called the first fundamental theorem of Monte Carlo for it gives an estimation of the chances of generating a large deviation in a Monte Carlo calculation. For definiteness let \( \delta = \frac{1}{100} \). Then the inequality becomes

\[
P\{ (G - \langle G \rangle)^2 \geq 100 \text{var}\{G\} \} \leq \frac{1}{100}
\]

or, using 2.38 when \( \text{var}\{G\} = (1/N)\text{var}\{g\} \),

\[
P\left( (G - \langle G \rangle)^2 \geq \frac{100}{N} \text{var}\{g\} \right) \leq \frac{1}{100}.
\]

Since by making \( N \) big enough, the variance of \( G \) becomes as small as one likes, the probability of getting a large deviation relative to \( \delta \) between the estimate of the integral and the actual value becomes very small. For large sample size (large \( N \)), the range of values of \( G \) that will be observed with some fixed probability will be contained in a region of decreasing size near \( \langle g \rangle \). This is the heart of the Monte Carlo method for evaluating integrals.
A much stronger statement than the Chebychev inequality about the range of values of $G$ that can be observed is given by the central limit theorem of probability. For any fixed value of $N$, there is a pdf that describes the values of $G$ that occur in the course of a Monte Carlo calculation.* As $N \to \infty$, however, the central limit theorem shows that there is a specific limit distribution for the observed values of $G$, namely, the normal distribution (see Section 2.4). Set

$$G_N = \frac{1}{N} \sum_{i=1}^{N} g(X_i)$$

and

$$t_N = \frac{(G_N - \langle g(X) \rangle)}{\sqrt{\text{var}(G_N)}}$$

then

$$\lim_{N \to \infty} P\{a < t_N < b\} = \int_{a}^{b} \frac{\exp[-t^2/2]}{\sqrt{2\pi}} \, dt.$$ (2.41)

Let $\sigma^2 = \text{var}(g)$. Eq. 2.41 can be rewritten so as to specify a probability distribution function for values of $G_N$:

$$f(G_N) = \frac{1}{\sqrt{2\pi(\sigma^2/N)}} \exp \left[ -\frac{N(G_N - \langle g \rangle)^2}{2\sigma^2} \right].$$

As $N \to \infty$, the observed $G_N$ turns up in ever narrower intervals near $\langle g \rangle$ and one can predict the probability of deviations measured in units of $\sigma$. That is, the observed $G_N$ is within one standard error (i.e., $\sigma/\sqrt{N}$) of $\langle g \rangle$ 68.3% of the time, within two standard errors of $\langle g \rangle$ 95.4% of the time, and within three standard errors 99.7% of the time.

The central limit theorem is very powerful in that it gives a specific distribution for the values of $G_N$, but it applies only asymptotically. How large $N$ must be before the central limit theorem applies depends on the problem. If for a particular problem the third central moment $\mu_3$ (Eq. 2.7) of $g$ exists, then the central limit theorem will be substantially satisfied when

$$|\mu_3| << \sigma^3 \sqrt{N}.$$  

Then confidence limits derived from the normal distribution can be applied to the results of a Monte Carlo calculation.

* For fixed $N$, if $G$ was calculated $M$ times, each time with a different sequence of iid random variables, the set $\{G_j\}, \ j = 1, \ldots, M$ has a specific distribution function.
2.7 Distribution of the mean of a random variable: A fundamental theorem

Tab. 2.1  Confidence limits for the exponential distribution as $N$ increases

| $N$ | $\sqrt{N}$ | $\overline{x} < \mu - 2\sigma$ | $\overline{x} < \mu - \sigma$ | $\overline{x} > \mu + \sigma$ | $\overline{x} > \mu + 2\sigma$ | $|\overline{x} - \mu| > 2\sigma$ |
|-----|------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-----------------------------|
| 5   | 2.24       | 0.00022                       | 0.1468                        | 0.1525                        | 0.04098                       | 0.04120                     |
| 10  | 3.16       | 0.00463                       | 0.1534                        | 0.1554                        | 0.03685                       | 0.04148                     |
| 20  | 4.47       | 0.00977                       | 0.1563                        | 0.1569                        | 0.03335                       | 0.04312                     |
| 50  | 7.07       | 0.01466                       | 0.1578                        | 0.1579                        | 0.02981                       | 0.04447                     |
| 100 | 10.0       | 0.01711                       | 0.1582                        | 0.1583                        | 0.02786                       | 0.04497                     |
| 200 | 14.14      | 0.01881                       | 0.1584                        | 0.1585                        | 0.02643                       | 0.04524                     |
| 500 | 22.36      | 0.02028                       | 0.1586                        | 0.1586                        | 0.02511                       | 0.04539                     |
|     | Normal     | 0.02275                       | 0.1587                        | 0.1587                        | 0.02275                       | 0.04550                     |

It is possible to calculate confidence limits exactly for the average of $N$ variates drawn from certain distributions, such as the exponential distribution, $\exp(-x), 0 \leq x < \infty$. For this distribution, $\sigma = 1; \mu_3 = 2$. Table 2.1 gives the chance that the mean $\overline{x}$ of $N$ random variates drawn from this distribution differs from $\mu$ by one or two times $\sigma$ on either side. Since the exponential distribution is skewed—there are no negative values of $x$—there are more large values of the average than predicted from the normal distribution. The last column shows that the total probability for deviations more than two standard deviations, large and small, converges quickly to that given by the normal distribution. Whether this distinction matters will depend upon the application.

Without the central limit theorem, there is in general only the much weaker upper bound of the Chebychev inequality to suggest how much the observed $G_N$ deviates from the actual mean. Of course in specific cases, studies can be made of the distribution of the estimator. Much Monte Carlo is done assuming that the theorem has been satisfied no matter what the sample size; reported errors must be considered optimistic in such cases.

When the variance is infinite it is sometimes possible to find a limit distribution for $G$ that will lead to a central limit theorem for that particular problem. The limit distribution will in general not be normal. The Cauchy distribution yields an elementary example, as will be seen later.

The variance used in the discussion given above may itself be estimated using independent values of $g(x_n)$ in the following way:

$$\left(\frac{1}{N} \sum_{i=1}^{N} g^2(X_i) - \left[ \frac{1}{N} \sum_{i=1}^{N} g(X_i) \right]^2 \right)^2 = \langle g^2 \rangle - \frac{1}{N^2} \left( \sum_{i=1}^{N} g(X_i)^2 + \sum_{i,j \neq j=1}^{N} g(X_i)g(X_j) \right).$$

(2.42)

Using the independence of $g(X_i)$ and $g(X_j)$ in evaluating $\langle g(X_i)g(X_j) \rangle$, we find the right-hand side equal to

$$\left(1 - \frac{1}{N} \right) \langle g^2 \rangle - \frac{N(N-1)}{N^2} \langle g \rangle^2 = \frac{N-1}{N} \text{var}\{g\}.$$
Thus an estimator for $\sigma ^2$ is

$$
\sigma ^2 \approx \frac{N}{N-1} \left\{ \frac{1}{N} \sum _{i=1} ^N g^2(X_i) - \left( \frac{1}{N} \sum _{i=1} ^N g(X_i) \right)^2 \right\} .
$$

(2.43)

An estimator of the variance of the estimated mean is given by

$$
\text{var}\{G_N\} \approx \frac{1}{N-1} \left\{ \frac{1}{N} \sum _{i=1} ^N g^2(X_i) - \left( \frac{1}{N} \sum _{i=1} ^N g(X_i) \right)^2 \right\} .
$$

(2.44)

2.8

Distribution of sums of independent random variables

Let $X$ be chosen from $f_1(x)$ and $Y$ independently chosen from $f_2(y)$. If the sum $Z = X + Y$ is formed, what is the probability distribution function for $Z$?

The distribution function is defined as

$$
F_3(z) = P\{Z \leq z\} = P\{X + Y \leq z\}.
$$

Since $X$ and $Y$ are independent, their joint probability distribution function is

$$
f(x,y) = f_1(x)f_2(y).
$$

The variables $x$ and $y$ can be considered to form a point in the $xy$ plane (see Figure 2.4). What fraction of the time does the point $(x, y)$ lie below the line $X + Y = Z$? The (cumulative) distribution function is

$$
F_3(z) = \int _{x+y\leq z} \int f_1(x)f_2(y) \, dx \, dy
$$

$$
= \int _{-\infty} ^z \int _{-\infty} ^{z-x} f_2(y)f_1(x) \, dy \, dx.
$$

The cumulative distribution function of $y$ is

$$
F_2(y) = \int _{-\infty} ^y f_2(u) \, du,
$$

so

$$
F_3(z) = \int _{-\infty} ^z F_2(z-x)f_1(x) \, dx.
$$

(2.45)

Differentiating with respect to $z$, one obtains for the pdf of $z$

$$
f_3(z) = \int _{-\infty} ^z f_2(z-x)f_1(x) \, dx .
$$
This is a convolution and Fourier transforms can be used to evaluate it. Let

\[
c_1(t) = \int_{-\infty}^{\infty} e^{ixt} f_1(x) \, dx = E(e^{itX}). \tag{2.46}
\]

In probability theory, \( c_1(t) \) is labeled the characteristic function of \( X \). The characteristic function of the sum \((X + Y)\) is

\[
c_3(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp[i(x+y)t] f_1(x) f_2(y) \, dx \, dy = c_1(t)c_2(t), \tag{2.47}
\]

or the characteristic function of a sum is the product of the characteristic functions of the terms of the sum.\(^*\) Clearly induction gives the same result for a sum of \( n \) variables. The characteristic function may be inverted (by a Fourier transform) to give the pdf to which it corresponds. If \( n \) identical functions constitute the sum, this result can be used to prove the central limit theorem.

When \( F(x) \) is the normal distribution (see Table 2.1), \( \phi(x|0,1) \), the characteristic function is \( \exp[-t^2/2] \). If a normal distribution is sampled \( n \) times,

\(^*\) Note that if the two random variables are not independent, this statement is generally not true.
the sum is also distributed normally. The characteristic function is then 
\[ [c(t)]^n = \exp[-nt^2/2], \] which when inverted gives the normal distribution 
\( \phi(x;0,n^{1/2}) \). A similar conclusion follows for the Cauchy distribution; the 
characteristic function is \( \exp[-|t|] \), and after \( n \) samples, the characteristic 
function of the sum is \( [c(t)]^n = \exp[-n|t|] \). The distribution of the sum of 
\( n \) Cauchy variables has a Cauchy distribution and the "width" of this Cauchy 
distribution increases as \( n \). As a final example consider the exponential distri-
bution \( \lambda \exp[-\lambda t] \); its characteristic function is 
\[ c(t) = \frac{1}{1 - it/\lambda}. \]

Therefore a sum of exponential random variables will not be distributed exponentially.

Note that the width of the distribution of the sum of \( n \) random variables in-
creases with \( n \). This is not contradictory with the earlier result on the mean of 
\( n \) random variables. It is not difficult to show—we shall prove it later—that 
if the characteristic distribution for \( x \) is \( c(t) \), then the characteristic function \( x/n \) is \( c(t/n) \). From this it follows that if \( n \) variables are drawn from 
the normal distribution \( \phi(x;0,1) \), then the mean has characteristic function 
\( \exp[-(n/2)(t/n)^2] = \exp[-t^2/2n] \). This may be inverted to give the distribu-
tion function \( \phi(x;0,n^{-1/2}) \) for the mean. The latter shrinks with \( n \) as predicted 
by the central limit theorem. Indeed, we see that here the limiting behavior holds exactly for any \( n \).

In the same way the characteristic function for the mean of \( n \) Cauchy vari-
bables is \( \exp[-n|t|/n] = \exp[-|t|] \). Thus the distribution of the mean is the 
same as for a single variable. Again the limit distribution for the mean is exact 
for any \( n \), but now the distribution does not change its width. This does not, 
of course, satisfy the central limit theorem because the Cauchy distribution 
does not satisfy the requirement of a finite variance.

### 2.9 Monte Carlo integration

We may summarize the important results of this chapter as follows. If 
\( X_1, X_2, \ldots, X_n \) are iid random variables with probability distribution function 
\( f(x) \) (\( x \) does not necessarily have to be in \( \mathbb{R}^1 \), \textit{i.e.} be on the real line), then for 
a function \( g(x) \), an estimator is 
\[ G_N = \frac{1}{N} \sum_{i=1}^{N} g(X_i), \]

\[ \langle G_N \rangle = \int_{-\infty}^{\infty} f(x)g(x) \, dx, \]
and
\[ \text{var} \{ G_N \} = \frac{1}{N} \text{var} \{ g \}. \]

As \( N \to \infty \) and if the variance exists, the distribution of possible values of \( G_N \) narrows about the mean as \( N^{-1/2} \); or the probability of finding a \( G_N \) some fixed distance away from \( \langle G_N \rangle \) becomes smaller.

In the development of the Monte Carlo method so far, it has been assumed that the random variables are drawn from a continuous distribution function and that they are used to approximate an integral. Similar procedures can be employed to perform sums by Monte Carlo. It becomes advantageous to use Monte Carlo methods in the discrete case when many indices are involved. Consider the sum \( \sum_i p_i g(X_i) \), where \( p_i \) is a discrete probability distribution. If random variables \( X_1, \ldots, X_M \) are sampled from \( p_i \) and the quantity
\[ G = \frac{1}{M} \sum_{i=1}^{M} g(X_i) \]
formed, the expected value of \( G \) is an estimator for the sum
\[ \langle G \rangle = \sum p_i g(X_i). \]

Monte Carlo evaluation of a sum might be used to determine the probability of winning at solitaire; the probability is a finite sum but contains a large number of terms. A sum over the permutations of \( L \) objects becomes cumbersome when \( L \) is large, but a Monte Carlo calculation can be performed efficiently.

The basic random variable used in Monte Carlo has been set by historical convention to be distributed uniformly between 0 and 1.

\[ f_u(x) = 1, \quad 0 \leq x \leq 1. \]

The use of this random variable in integration is demonstrated in the following example:
\[ \int_0^1 \sqrt{1 - x^2} \, dx = \frac{\pi}{4}. \] (2.48)

It can be rewritten
\[ \int_0^1 f_u(x) \sqrt{1 - x^2} \, dx = \frac{\pi}{4}, \]

The integral is now in a form in which we can apply the method described above for evaluating integrals. A uniform random variable \( \xi \) is sampled from \( f_u(x) \), \( g(\xi) = \sqrt{1 - \xi^2} \) is calculated and this process is repeated \( N \) times to form
\[ G_1 = \frac{1}{N} \sum g(\xi_i) = \frac{1}{N} \sum (1 - \xi_i^2)^{1/2}. \] (2.49)
The algorithm for calculating $G_1$ can be easily programmed. In the pseudocode that follows we shall use rand(u) as a generic name for a utility that produces good pseudorandom numbers uniform on (0, 1) and assume that succeeding values returned by rand(u) are independent.

Initialize sum=0
Repeat N times, i = 1, . . ., N
    sum = sum + sqrt(1.0 - rand(u)^2)
EndRepeat
mean = sum/N

However, the evaluation of the integral in Eq. 2.48 by Monte Carlo can be approached in an entirely different manner.

Consider the unit square in the xy plane and the circle with unit radius (Figure 2.5). Integrating over the unit square but counting only those pairs of x and y that lie within the quarter circle yields the area of the quarter circle. That is

$$\int_0^1 \int_0^1 f(x, y)g(x, y) \, dx \, dy = \frac{\pi}{4},$$

where

$$f(x, y) = \begin{cases} 1 & (x, y) \text{ in } (0, 1) \otimes (0, 1) \text{ (i.e., inside the unit square)} \\ 0 & \text{otherwise} \end{cases}$$

and

$$g(x, y) = \begin{cases} 1, & x^2 + y^2 \leq 1 \\ 0, & x^2 + y^2 > 1. \end{cases}$$

Since x and y are independent,

$$f(x, y) = f_u(x)f_u(y),$$
so that \( f(x, y) \) may be sampled by drawing two independent uniform random variables, \( \xi_i, \eta_i \) and forming the estimator, \( G_2 \), which sums up the number of times \( \xi_i^2 + \eta_i^2 \leq 1 \). An algorithm for performing the Monte Carlo integration in pseudocode is,

Initialize sum = 0
Repeat \( N \) times, \( i = 1, \ldots, N \)
  \( x = \text{rand}(u) \)
  \( y = \text{rand}(u) \)
  If \( x^2 + y^2 \leq 1.0 \) then
    sum = sum + 1
  EndIf
EndRepeat
mean = sum/N

The two methods are of comparable effectiveness. Of course, neither is a method of choice for this (or other one-dimensional integrals.) Which of the two estimators, \( G_1 \) or \( G_2 \), will be used in actual practice depends on factors such as the variance and efficiency of the calculation.

But a seemingly paradoxical outcome is seen: an integral in one dimension is recast as an integral in two without serious, if any, loss of computational efficiency. That one can work in many dimensions—indeed one can add extra dimensions— is a characteristic of Monte Carlo quadrature in contrast to discrete numerical quadrature, and is a property that can be exploited to great advantage in many applications.

The procedure detailed above may easily be generalized. To evaluate the \( L \)-dimensional integral over the unit hypercube

\[
\int \cdots \int g(x_1, x_2, \ldots, x_L) \, dx_1 \, dx_2 \cdots dx_L,
\]

\( L \) uniform random variables could be sampled, the function \( g(x_1, x_2, \ldots, x_L) \) calculated, and the whole process repeated \( N \) times. The arithmetic mean of the function values gives an unbiased estimate of the integral.

Of course, given the ability to generate random variables from any distribution over any space (as discussed in the next chapter) the domain of integration need not be limited to hypercubes.

When a variable is generated using \( \text{rand}(u) \) or a similar function, it is a pseudorandom variable; that is, it was generated by a deterministic algorithm. Since it is possible to use truly random variables in any calculation, why are pseudorandom variables used? An absolute requirement in debugging a computer code is the ability to repeat a particular run of the program. If truly random numbers were used, an identical calculation could not be repeated and the recurrence of an error would be left to chance. It has been suggested that
pseudorandom numbers be used to debug a program and that in the actual exercising of the program truly random numbers be used. This method also suffers from the inability to repeat a particular calculation. If after many hours of computer time an error should occur in a code (subtle logical errors occur in many Monte Carlo codes), it is of utmost importance to be able to repeat the error at will as an aid to debugging. It is also very useful to be able to repeat a calculation when changes are made or when the program is moved to a different computer. Furthermore, lengthy calculations may require the generation of an enormous number of “random variables” at great rates, and deterministic algorithms can be created to fill this need. The generation of pseudorandom variables is discussed in much greater detail in Chapter 9.

Finally, there are situations in which the same sequence of random variables must be generated, for example, to change a parameter but retain the correlation with a previous calculation. In that case, the use of pseudorandom sequences offers a great advantage.

2.10 Monte Carlo Estimators

We have defined an estimator as a useful approximation to a quantity of interest \(Q\), which may be derived from a Monte Carlo calculation. In the example given in Eqs. 2.48 and 2.49 in which

\[
\frac{\pi}{4} = \int \sqrt{1 - x^2} \, dx \approx \frac{1}{N} \sum_i \sqrt{1 - \xi^2_i},
\]

we consider \(Q\) to be \(\pi/4\) and our estimator is the approximation on the right-hand side, which is a function \(\theta(\xi_1, \xi_2, \ldots, \xi_N)\) of the \(N\) random or pseudorandom variables used in the calculation.

The function \(\theta\) is of course itself random, and the statement that it gives a satisfactory approximation to \(Q\) means that it is not expected to fluctuate far from \(Q\). Put a little more formally,

\[
\langle (\theta - Q)^2 \rangle / Q^2 \ll 1.
\]

Acceptable values of the ratio depend on the application. A Monte Carlo calculation may be intended to give a rough estimate of some numerical quantity, or it may be aimed at high precision, or at a target in between. The appropriate balance between small or zero bias and small variance will depend on these choices.

We write

\[
\langle (\theta - Q)^2 \rangle = \langle (\theta - \langle \theta \rangle)^2 \rangle + \langle (\theta) - Q \rangle^2.
\]
and observe that the quality of $\theta$ as a measure of $Q$ comes separately from the variance of $\theta$ and from the departure of its mean from $Q$. The quantity $\langle \theta \rangle - Q$ is called the bias of the estimator. An unbiased estimator is one for which $\langle \theta \rangle = Q$ for any experiment whatever the number $N$ may be.

The quadratures we have discussed are unbiased since the result is linear in the functions calculated. For some problems, however, it is very difficult to formulate unbiased estimators. As we shall see, there are many problems for which the answer required is a ratio of integrals,

$$Q = \frac{\int_0^1 g_1(x) \, dx}{\int_0^1 g_2(x) \, dx},$$

for which a suitable estimator is

$$\theta(\xi_1, \ldots, \xi_N) = \frac{\sum_i g_1(\xi_i)}{\sum_i g_2(\xi_i)}.$$

Since this is not a linear function of $g_2$, it is biased. An example that can easily be analyzed is

$$1 = \frac{1}{\int_0^\infty xe^{-x} \, dx} \approx \frac{N}{\sum_{i=1}^N \log (\xi_i)}.$$

The random variable $x$ is sampled from an exponential distribution (see Section 3.1) and an estimator for the quotient is formed. It can then be shown that

$$\left\langle \frac{N}{\sum_{i=1}^N \log (\xi_i)} \right\rangle = \frac{N}{N-1} \to 1 + \frac{1}{N} + \frac{1}{N^2} + \cdots.$$

Our estimator is biased by $1/N$, which, for large $N$, decreases faster than the standard error of the mean $(\sigma / \sqrt{N})$, where here $\sigma = 1$. This $1/N$ behavior is typical of the bias of such ratios. The results that may be derived from a Monte Carlo calculation are more general than this, and may have different variation of the bias. It is of course best if the bias becomes 0 as $N$ grows large.

An estimator $\theta$ is termed consistent for the quantity $Q$ if $\theta$ converges to $Q$ with probability 1 as $N$ approaches infinity. That is, $\theta$ is a consistent estimator of $Q$ if

$$P\left\{ \lim_{N \to \infty} \theta(\xi_1, \xi_2, \ldots, \xi_N) = Q \right\} = 1.$$

The law of large numbers states that the sample mean $\bar{X}_N$ is a consistent (and unbiased) estimator of the mean $\mu$. It further implies that estimators of quotients that are quotients of means are also consistent (although, in general, biased).
While unbiased estimators are desirable, they should not be introduced at
the expense of a large variance, since the overall quality is a combination of
both bias and consistency. In general one seeks the minimum of 
\[ \ell(q) \]

An example may clarify the issue. Suppose \( \xi_i \) are drawn uniformly and
independently on the interval \((0, 1)\) and we wish to estimate the mean, \( Q \),
from the chosen \( \xi_i \). The estimator

\[ X_N = \theta_1 = \frac{1}{N} \sum \xi_i. \]
is the usual one discussed above. A plausible alternative for large \( N \) is

\[ \theta_2 = \frac{1}{2} \max(\xi_1, \xi_2, \ldots, \xi_N). \]

Note that this estimator always gives results less than the mean, \( \frac{1}{2} \). It is easy
to show that

\[
E(\theta_1) = Q = \frac{1}{2}, \\
\text{var}\{\theta_1\} = \frac{1}{12N} = O\left(\frac{1}{N}\right),
\]

while

\[
E(\theta_2) = \frac{N}{N+1} Q = Q(1 + O\left(\frac{1}{N}\right)), \\
E(\theta_2 - Q)^2 = \frac{2Q^2}{(N+1)(N+2)} = O\left(\frac{2}{N^2}\right).
\]

Thus, although \( \theta_2 \) is biased (by \( 2/N \)), its variance for large \( N \) is smaller than
that of \( \theta_1 \) by a ratio of \( N/3 \). The bias is then equal to its standard deviation.
For some purposes, this would be a more useful estimator.

Just as a good Monte Carlo calculation must be supplemented with an es-
timate of the statistical error, sources of bias should be identified. The bias
should be estimated numerically or an upper bound should be determined. A
useful way of estimating bias when the behavior with \( N \) is known is to group
the data in samples smaller than \( N \), say \( n = N/m \). One can average this more
biased estimator over the \( m \) groups obtained and study the dependence on \( m \):

\[
\text{bias of } \frac{\sum_{i=1}^{\ell_1} g_1(\xi_i)}{\sum_{i=1}^{\ell_2} g_2(\xi_i)} \approx \frac{c}{N}, \\
\text{bias of } \frac{1}{m} \sum_{l=1}^{m} \left( \frac{\sum_{i=n(l-1)+1}^{n(l)} g_1(\xi_i)}{\sum_{i=n(l-1)+1}^{n(l)} g_2(\xi_i)} \right)_{\text{group}_l} = \frac{c}{n} = \frac{cm}{N},
\]

with \( c \) a positive constant.
We note in passing that this method of grouping is also a practical way of estimating the variance of the quotient. This consists in selecting groups of numerators and denominators that are nearly independent, forming partial quotients for the groups, and then applying Eq. 2.44.
Bibliography

References


General references for further reading

Elementary


More Advanced