Some Probability Distributions and Their Uses

The range of all possible probability experiments covers a lot of ground, from the outcome of a weighted coin toss to next month’s price of a given stock, from the arrival time of the next phone call at a switchboard to the birth weight of a mother’s first child. To simulate these and other probability experiments we need to know how to convert uniformly distributed random numbers into a sample from whatever distribution underlies the experiment.

In this chapter we introduce some of the major probability distributions and show how to use the computer to draw a random sample from each of them. This means converting samples $U$ drawn from the uniform $[0,1)$ distribution $U(0,1)$ into samples from the required density. We will present the probability density and cumulative distribution functions for each distribution along with their means and variances. We will show how the random samples can be used to calculate probabilities and solve problems.

The main techniques used for sampling are *cdf inversion, simulation, composition, mapping*, and *rejection*. Our objective is to introduce these sampling methodologies and to illustrate them. In many cases, better, but more involved, methods are available for sampling these distributions more efficiently should timing be a critical issue. See for example [Rip87, BFS83, Rub81].

2.1 CDF Inversion–Discrete Case: Bernoulli Trials

Suppose an experiment has just two outcomes, say “failure” and “success” (or off and on, or 0 and 1, etc.). Such an experiment is called a *Bernoulli trial*. A coin flip is an example of a Bernoulli trial, and so
is a roulette wheel with just two sectors. Let \( p \) be the probability of a success and \( q = 1 - p \) the probability of failure. Then the pdf of a Bernoulli trial is (using 0 for failure and 1 for success)

\[
f(x) = \begin{cases} 
1 - p, & \text{if } x = 0, \\
p, & \text{if } x = 1, 
\end{cases}
\]

\[= p^x(1-p)^{1-x}, \quad x = 0, 1.\]

The cdf is

\[
F(x) = \begin{cases} 
0, & \text{if } x < 0, \\
1 - p, & \text{if } 0 \leq x < 1, \\
1, & \text{if } 1 \leq x. 
\end{cases}
\]

Consequently, using (1.10) and (1.11), the mean of a Bernoulli trial is

\[\mu = 0 \cdot (1 - p) + 1 \cdot p = p\]

and the variance is

\[\text{var} = (0 - p)^2(1 - p) + (1 - p)^2 p = p(1 - p)(p + 1 - p) = pq.\]

### 2.1.1 Two-Outcome CDF Inversion

Drawing a Bernoulli trial sample \( T \) using \( U \sim U(0, 1) \) is very easy. If \( U < p \) then declare that a “success” has occurred; otherwise, declare a “failure.” This works because the \( U(0, 1) \) density is pdf\((t) = 1 \) for \( 0 \leq t \leq 1 \), and so

\[
\Pr(T = \text{success}) = \Pr(U < p) = \int_0^p \text{pdf}(t) \, dt = p.
\]

For example, suppose we want to simulate the spin of a casino roulette wheel and see whether the outcome is red. There are 18 red and 18 black sectors in a casino roulette wheel and two green sectors, 0 and 00. Therefore the success probability is \( p = 18/38 = 0.4736 \ldots \). Using the random number generator we generate \( U \) uniformly at random in \([0,1]\) and declare that “red” occurred if \( U < 0.4736 \ldots \); otherwise, “not red” occurred.

Alternatively, since \( q = 1 - p = 0.5263 \ldots \), we could test whether \( U < q \). In this case our sample is “not red” if so and “red” if \( U \geq 0.5263 \ldots \). This is not a better method than before, but it is consistent with the cumulative distribution function that is shown in Figure 2.1. The interpretation is that \( U \) will lie between 0 and 1 on the \( y \)-axis. If \( U \) is less than \( q = 1 - p \) return 0, meaning “not red” here, and if \( U \) equals or exceeds \( 1 - p \), return 1, meaning “red.”
2.1 CDF Inversion–Discrete Case: Bernoulli Trials

2.1.2 Multiple-Outcome Distributions

The two-outcome Bernoulli trial is the simplest discrete distribution. More generally, discrete distributions have several possible outcomes. If their number is not too large, we can use the cdf graph to draw samples with the correct frequencies as hinted at above. Some examples will make this clear.

Suppose we want to simulate the number of eggs laid by a shorebird during breeding season. These birds have $X = 2, 3, 4,$ or rarely $5$ eggs per clutch. Suppose the observed frequencies of these sizes are $p_2 = 0.15, p_3 = 0.20, p_4 = 0.60,$ and $p_5 = 0.05$ respectively. The cdf of $X$ is shown in Figure 2.2.

Let $U \sim U(0, 1)$ be a uniform $[0, 1)$ sample. Starting from the point $(0, U)$ (on the $y$-axis), proceed to the right until you encounter a jump in the cdf, a vertical dashed line segment. Now proceed down to the $x$-axis, and return this value as the selection. As indicated along the $y$-axis, “2” is selected with 0.15 probability, since its interval occupies this fraction of the unit interval. Similarly, “3” is selected with 0.2 probability, since its interval occupies this fraction of the unit interval, and so on.

This technique is called inverting the cdf. For a discrete distribution it can be implemented on the computer in several ways.

The most general method sets up breakpoints in the interval $[0, 1]$ and then determines into which subinterval $U \sim U(0, 1)$ falls. The following code will generate one sample from $X$:
Matlab

```matlab
>> % set up the outcomes
>> oc(1)=2; oc(2)=3; oc(3)=4; oc(4)=5;
>> % set up the breakpoints
>> bp(1)=.15; bp(2)=bp(1)+.2; bp(3)=bp(2)+.60; bp(4)=1;
>> % end of set up, go here to get samples
>> U=rand; k=1;
>> while( U >= bp(k))
>>     k=k+1;
>> end % k is selected
>> selected = oc(k)
```

A downside of this method is that the program must step through the `while` loop for each point returned. The search can be sped up by reordering the outcomes according to decreasing probability; this reduces the expected time in the `while` loop. For example, put the first breakpoint at 0.6 mapping to 4 eggs, then 0.6 + 0.2 = 0.8 for 3 eggs and so on. In the first ordering, the expected number of times through the loop is

$$(1)(0.15) + (2)(0.2) + (3)(0.6) + (4)(0.05) = 2.55,$$

while in the second ordering the expected number of times through the loop is

$$(1)(0.6) + (2)(0.2) + (3)(0.15) + (4)(0.05) = 1.65.$$

While this may not seem like a great advantage, if thousands or millions of selections are needed it adds up.

As a much faster alternative, if the probabilities are simple two-place decimal fractions (as in this example), we can segment an array...
of length 100 into all the alternatives and then sample uniformly from this array. Here is the program.

```matlab
> i=1;
> for j=1:15
  oc(i)=2; i=i+1;
end;
> for j=1:20
  oc(i)=3; i=i+1;
end;
> for j=1:60
  oc(i)=4; i=i+1;
end;
> for j=1:5
  oc(i)=5; i=i+1;
end;
> k=floor(100*rand)+1;
> selected = oc(k)
```

Finally, the fastest way to generate a large number of samples is the vector method. The following code will generate 10,000 samples from $U(0, 1)$.

```matlab
> b(1)=.15; b(2)=b(1)+.2; b(3)=b(2)+.6;
> U=rand(1,10000); %10000 U(0,1) samples
> w2 = (U<b(1))*2;
> w3 = (U>=b(1) & U<b(2))*3;
> w4 = (U>=b(2) & U<b(3))*4;
> w5 = (U>=b(3))*5;
> w=w2+w3+w4+w5; % w a random 10000
> % vector with the right frequencies
```

## 2.2 Walker’s Alias Method: Roulette Wheel Selection

Multiple outcome distributions with arbitrary assignments of probabilities are like a (generalized) roulette wheel. It is no loss of generality to assume that the finite outcomes are labeled $1, 2, \ldots, n$. Assume that their corresponding probabilities are $p_1, p_2, \ldots, p_n$. Selecting an outcome is like spinning a roulette wheel with sectors having central angle $p_12\pi$ for outcome 1, $p_22\pi$ for outcome 2, \ldots, and $p_n2\pi$ for outcome $n$.

The alias method takes more time to set up than cdf inversion, but runs faster because there is no search loop. The method uses two arrays,
an array of probability adjustments $Q(\cdot)$ and an array of aliases $A(\cdot)$. Again we demonstrate the method by means of an example.

In some species the female can affect the fraction of males versus females among her offspring depending on various factors. For example, the rufous vanga female produces more male offspring if she mates with a large male but produces more female offspring if she mates with a small male (see [Asa05]). The rationale is that a large male will beget larger offspring, and large males mate more often and more widely than small males. So if she mates with a large male, she can more effectively propagate her genes by having males. Since females will always be able to mate, if she has small offspring, it is better to have females.

Suppose we want to sample the number $X$ of female offspring in a clutch of size 8 given the statistics shown in Table 2.1. Since there are $M = 9$ possible outcomes, letting $Y = \text{int}[UM]$, the integer part of the product $UM$, where $U \sim U(0,1)$, selects 0, 1, 2, \ldots, 8, equally likely, at about 11%. Therefore we use the $Q$ array to adjust this down for those probabilities that need to be lower. Meanwhile, the excess probability is sent to some index $A(Y)$ whose selection probability exceeds 11%.

<table>
<thead>
<tr>
<th># female offspring</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>0.02</td>
<td>0.04</td>
<td>0.08</td>
<td>0.15</td>
<td>0.2</td>
<td>0.21</td>
<td>0.16</td>
<td>0.09</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 2.1. Probability for female offspring in a clutch of 8.

The method works like this.

**Alias Algorithm**

1. let $U \sim U(0,1)$ and put $Y = \text{int}[UM]$ and $Z = \text{frac}(UM)$, where $\text{frac}(x) = x - \text{int}[x]$ is the fractional part of $x$.
2. if $Q(Y)/M < Z$ return $X = Y$, else return $X = A(Y)$.

Arrays that work for our example problem are shown below:

<table>
<thead>
<tr>
<th>$Q$</th>
<th>0.18</th>
<th>0.36</th>
<th>0.72</th>
<th>1.0</th>
<th>0.97</th>
<th>0.43</th>
<th>0.68</th>
<th>0.81</th>
<th>0.45</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>-</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

To see why this works, first note that $Y$ is an equally likely random variable on $0, 1, 2, \ldots, M-1$, and $Z$ is independent of $Y$ and distributed as $U(0,1)$ (the argument that $Y$ and $Z$ are independent is that $[0, 1)$ is partitioned into $M$ equal parts, with $Y$ selecting one of these parts and $Z$ indicating where in this subinterval $U$ is located). Therefore the if
clause in Step 2 of the algorithm succeeds with probability \( Q(Y)/M \) and selects \( Y \). Otherwise, \( A(Y) \) is selected with probability \((1 - Q(Y))/M\).

Since the roulette wheel for \( X = 0 \) is smallest, being 0.02, we start with it. Solving \( Q(0)/M = 0.02 \) gives \( Q(0) = 0.18 \), making index 0 occur with the correct frequency. We send the excess probability, amounting to \((1 - Q(0))/M = 0.091\), to the index with maximum roulette wheel sector, namely 5, so \( A(0) = 5 \). Now 9.1% of index 5’s probability is taken care of, so it needs only \( 0.21 - 0.091 = 0.209 \) more. We continue in this way using \( Q \) to adjust the uniform probability \( 1/M \) downward as needed and sending the excess probability to whatever index still needs more than \( 1/M = 0.11 \). If we send too much to some index, that causes no problem, since it too gets reduced by \( Q \) when that index’s \( Q \) is calculated.

The alias algorithm is an ingenious implementation of the composition method of sampling. We have decomposed the original roulette wheel density \( f \) into two components: the first, \( f_1 \), is \( Q(\cdot)/M \) with support on \( 0, 1, \ldots, 8 \) and the second, \( f_2 \), has support on the range of \( A \), here 3, 4, 5, 6. The fraction of the time that \( Q \) decides the outcome is the sum \( \sum_{i=0}^{8} Q(i)/9 = 0.622 \), and the remaining fraction of the time, namely 0.377, the outcome is decided by \( A \).

The composition method is a very general method and will be elaborated on shortly, in Section 2.8.

### 2.3 Probability Simulation: The Binomial Distribution

Often an entire series of independent Bernoulli trials is performed, and we are interested in the outcomes of all of them. Let the probability of a success be \( p \) and let \( q = 1 - p \). Let \( X \) stand for the number of successes in a series of \( n \) such trials. Then \( X \) is one of the values \( \Omega = \{0, 1, \ldots, n\} \), but they are not equally likely. Even in our simple experiment of tossing two coins, recall that the probability of obtaining 1 head is 1/2, while the probability of 0 or 2 heads is 1/4 each. This is because the outcome 1 Head can occur in two different ways, (H, T) or (T, H).

The number of ways in which there can be \( x \) successes in \( n \) trials is given by the quotient

\[
\frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)\cdots2\cdot1},
\]  

(2.1)
in which there are \( x \) factors in both the numerator and denominator. This is seen as follows. Suppose first that the successes are distinguished from each other by subscripts, \( S_1, S_2, \ldots, S_x \). Among the \( n \) trials there are \( n \) choices for where success \( S_1 \) occurs. To place \( S_2 \), one possibility has been taken (by \( S_1 \)), so there are \( n - 1 \) choices for where \( S_2 \) occurs. Continuing, we see that there are

\[
n(n - 1) \cdots (n - x + 1)
\]

(2.2)

ways of selecting on which trials the distinguished successes occur.

As an illustration of this, suppose the three symbols \( S \), \( s \), and \( F \) are used to make 4-letters words (we use upper- and lower-case \( S \) here instead of \( S_1 \) and \( S_2 \)). The rule is that one and only one \( S \) must be used and the same for \( s \). According to (2.2), there should be \( 4 \cdot 3 = 12 \) such words. They are (\( S \) in first place) \( SsFF \), \( SFsF \), \( SFFs \), (\( S \) in second place) \( sSFF \), \( FSsF \), \( FSFs \), (\( S \) in third place) \( sFSF \), \( FsSF \), \( FFSs \), and (\( S \) in fourth place) \( sFFS \), \( FsFS \), \( FFsS \).

Returning to the general case, in reality the success are not distinguished, so if we pick one of the possible choices (of distinguished placements), how many other choices are merely rearrangements of the successes? But this is just counting the number of different orders of the subscripted \( S \)'s, that is the permutations of \( x \) symbols. Using the same logic as above,

\[
\text{number of permutations of } x \text{ symbols } = x(x - 1)(x - 2) \cdots 2 \cdot 1 = x!.
\]

(2.3)

As indicated, the notation for this product is \( x! \), read \( x \) factorial. Hence the number of ways to obtain \( x \) successes in \( n \) trials is the quotient as predicted by (2.1). This ratio is denoted by \( \binom{n}{x} \), read \( n \) choose \( x \), or sometimes by \( C(n, x) \),

\[
C(n, x) = \binom{n}{x} = \frac{n(n - 1)(n - 2) \cdots (n - x + 1)}{x(x - 1) \cdots 2 \cdot 1} = \frac{n!}{x!(n - x)!}.
\]

(2.4)

It is easy to verify that the last member of this equation is an equivalent expression.

The quantity defined in equation (2.4) is also called a binomial coefficient, primarily because of its connection with the binomial theorem (given in equation (2.6)).

In our example, one of the words is \( SsFF \), but there are \( 2! = 2 \) ways of ordering \( S \) and \( s \), so this word pairs off with \( sFF \). Likewise,
SFsF pairs off with sFSF and so on. Hence, regarding S and s as not different, i.e., replacing s by S, there are six different possible words: SSFF, SFSF, SFFS, FSSF, FSFS, and FFSS. Similarly there are six different ways in which there can be two successes and two failures in four Bernoulli trials.

Now return to the problem of calculating the probability of \( x \) successes in \( n \) trials. We know that this can happen in \( \binom{n}{x} \) ways; pick any one of them, for example all successes in the first \( x \) trials and all failures in the last \( n - x \) trials. What is the probability of that? Since the trials are independent, it is the product \( p^x q^{n-x} \) (recall that \( p \) is the probability of success and \( q = 1 - p \) is the probability of failure). This is also the probability of each of the other possibilities. Hence the probability that there are \( x \) successes is given by

\[
\text{pdf}(x) = \Pr(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \ldots, n. \tag{2.5}
\]

For the discrete random variable \( X \), this is also its density function. This is the binomial distribution, often denoted by \( b(n, p) \), \( n \) independent Bernoulli trials with probability \( p \) of success on each trial.

Equation (2.5) is closely related to the binomial theorem for the expansion of the binomial \( (a + b)^n \),

\[
(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{2.6}
\]

In fact, summing (2.5) over \( x \) gives

\[
\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p + q)^n = 1,
\]

as expected, since \( \Pr(\Omega) = 1 \).

Mean and Variance

According to (1.10), the mean of the binomial distribution \( b(n, p) \) is

\[
\mu = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x} = np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}.
\]
Change variables by putting \( y = x - 1 \) and \( m = n - 1 \). Then \( n - x = m - y \), and we have

\[
\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x} = \sum_{y=0}^{m} \binom{m}{y} p^y q^{m-y} = (p+q)^m = 1.
\]

Therefore

\[
\mu = np. \tag{2.7}
\]

As we have already seen, expectation is linear; recall (1.12). Since the binomial random variable is the sum of \( n \) independent Bernoulli trials, it follows that the mean of the binomial distribution is the \( n \)-fold sum of the mean of a Bernoulli trial, just as asserted by (2.7).

Likewise, by Theorem 1.1 of Section 1.2.6, the variance of the binomial random variable is the sum of the variances of the \( n \) independent Bernoulli trials,

\[
\text{var} = npq. \tag{2.8}
\]

### 2.3.1 Sampling from the Binomial

If \( n \) is “small,” we could sample \( X \) using cdf inversion or the aliasing technique of the last section. Alternatively, and simpler to implement, a \( b(n,p) \) sample may be generated by simulation, that is, by \( n \) repetitions of the underlying Bernoulli trial.

For example, suppose we are interested in how many times red occurs in 20 consecutive plays of the casino roulette wheel. Find out by repeating the Bernoulli trial 20 times: sample \( U \sim U(0,1) \), check whether \( U < 18/38 \), and add 1 to the count if so.

If \( n \) is quite large, simulation is less desirable. Such is the case in the following problem. A mutation occurs to one of the daughter cells during cell division with probability \( p = 10^{-7} \). In one million cell divisions of a laboratory culture, how many mutations occur? In the next section we will see how to do this without drawing one million Bernoulli samples.

### 2.4 Another Simulation: The Poisson Distribution

Consider a problem in which the number of Bernoulli trials is very large indeed. This is the problem of counting the number of “successes” of a phenomenon when the successes occur at a known rate \( \lambda \). For example,
suppose telephone calls come into a 911 emergency switchboard randomly, but on the average, 1 every 10 minutes. How many calls will arrive between 10 and 10:30 A.M.?

In each second of time there is a very slight probability that there will be an incoming call, so the arrival or nonarrival of an incoming call in any given second is a Bernoulli trial. Thus in 30 minutes there are 1800 such Bernoulli trials.

To analyze this, we make four assumptions about the events of the phenomenon: the event rate $\lambda$ must be positive and

(a) The number of events occurring in nonoverlapping intervals are independent.
(b) The probability of exactly one event in a sufficiently short interval of length $h$ is approximately $\lambda h$.
(c) The probability of two or more events in a sufficiently short interval is negligible.

Let the random variable $X_t$ denote the number of events that occur in an interval of length $t$. Partition the interval into $n$ subintervals of equal length $t/n$, where $n$ is large enough that conditions (b) and (c) hold. Then whether an event occurs in a given interval is a Bernoulli trial with success probability $\lambda t/n$, and therefore the probability of $x$ events in the $n$ subintervals is

$$\binom{n}{x} \left( \frac{\lambda t}{n} \right)^x \left( 1 - \frac{\lambda t}{n} \right)^{n-x}.$$

Let $n \to \infty$ in this expression. Group the $x!$ in the denominator of $\binom{n}{x}$ with $(\lambda t)^x$, since $x$ is fixed. On the other hand, group $1/n^x$ with the remaining factors of the $\binom{n}{x}$ to get

$$\frac{n!}{n^x(n-x)!} = \frac{n(n-1) \cdots (n-x+1)}{n^x} \to_{n \to \infty} 1.$$

The remaining factor, $(1 - \lambda t/n)^{n-x}$, tends to $e^{-\lambda t}$ as follows

$$\lim_{n \to \infty} \left( 1 - \frac{\lambda t}{n} \right)^{n-x} = \lim_{n \to \infty} \left( 1 - \frac{\lambda t}{n} \right)^n \left( 1 - \frac{\lambda t}{n} \right)^{-x} = e^{-\lambda t} 1.$$  

Therefore

$$\Pr(X_t = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}. \quad (2.9)$$
This is the density for the Poisson distribution. Note that $\lambda t$ is a pure (dimensionless) number. If the event rate is $\lambda = 1/10$ min and we want answers about a 30-minute period, then $\lambda t = 3$.

The Poisson density also applies to “events” distributed in space as well as time. For example, suppose that knots occur in a certain batch of lumber as a Poisson process with a rate of 2 knots per eight foot stud on average. What is the probability of finding a stud having no knots? By (2.9), since $\lambda t = (2/8) \times 8$ here, it is

$$\Pr(X_{8} = 0) = \frac{(2)^0 e^{-2}}{0!} = e^{-2} = 0.135.$$ 

Note that $0!$ is 1 (by definition).

Since the mean of the binomial $b(n, \frac{\lambda t}{n})$ is $n \frac{\lambda t}{n} = \lambda t$ and is independent of $n$, it is natural to believe that the mean of the corresponding Poisson process is $\lambda t$. This is indeed the case. Figure 2.3 gives the details.

Recall the mutation problem of the previous section. The Bernoulli trial probability $p$ of a Poisson process is $\lambda t/n$, hence $\lambda t = np$. Therefore the Poisson approximation of this binomial is

$$\Pr(X = x) = \frac{(np)^x e^{-np}}{x!} = \frac{(0.1)^x e^{-0.1}}{x!}$$

for $n = 10^6$ and $p = 10^{-7}$. Compare this with (2.5). For $x = 0$ the exact probability is $(1 - p)^n = 0.904837331$ to nine decimal places, while the approximation is $e^{-0.1} = 0.904837418$.

Although the Poisson distribution can be used as an approximation to the binomial, we still have the problem of drawing samples from it. As we will see next, the technique for sampling from a Poisson distribution exactly is by simulation. But both the Poisson and binomial distributions for large $n$ can be sampled to a good approximation using the central limit theorem and the normal distribution covered below (see Section 2.6).

### 2.4.1 Sampling from the Poisson Distribution by Simulation

The following method will be explained in the exponential section to be taken up next. It is essentially a simulation technique, and is exact.

Let $U_0 = 1$ and $U_i$ be $U(0,1)$ samples for $i = 1, 2, \ldots$. Return the greatest integer $x$ such that
The calculation of the mean and variance of the Poisson distribution.

\[
\mu = \mathbb{E}(X(t)) = \sum_{k=0}^{\infty} k \frac{(\lambda t)^k}{k!} e^{-\lambda t} \\
= e^{-\lambda t} \left[ \frac{\lambda t}{0!} + \frac{(\lambda t)^2}{1!} + \frac{(\lambda t)^2}{2!} + \cdots \right] \\
= \lambda t e^{-\lambda t} e^{\lambda t} = \lambda t.
\]

And the variance of a Poisson random variable is var = \(\lambda t\):

\[
\text{var} = \mathbb{E}((X(t) - \mu)^2) = \sum_{k=0}^{\infty} (k - \lambda t)^2 \frac{(\lambda t)^k}{k!} e^{-\lambda t} \\
= e^{-\lambda t} \sum_{k=0}^{\infty} \left( k^2 - 2k\lambda t + (\lambda t)^2 \right) \frac{(\lambda t)^k}{k!} \\
= e^{-\lambda t} \sum_{k=0}^{\infty} \left( \frac{k^2}{k!} (\lambda t)^k - 2 \frac{k}{k!} (\lambda t)^{k+1} + \frac{(\lambda t)^{k+2}}{k!} \right) \\
= e^{-\lambda t} \left[ (\lambda t) \left( 1 + \frac{2}{1!} (\lambda t) + \frac{3}{2!} (\lambda t)^2 + \cdots \right) \\
- 2 \left( \frac{(\lambda t)^2}{0!} + \frac{(\lambda t)^3}{1!} + \cdots \right) + \left( \frac{(\lambda t)^2}{0!} + \frac{(\lambda t)^3}{1!} + \cdots \right) \right] \\
= e^{-\lambda t} \left[ (\lambda t) \frac{d}{d(\lambda t)} \left( \lambda t + \frac{(\lambda t)^2}{1!} + \frac{(\lambda t)^3}{2!} + \cdots \right) - 2(\lambda t)^2 e^{\lambda t} + (\lambda t)^2 e^{\lambda t} \right] \\
= (\lambda t) e^{-\lambda t} \frac{d}{d(\lambda t)} (\lambda t e^{\lambda t}) - (\lambda t)^2 \\
= (\lambda t) e^{-\lambda t} \left( e^{\lambda t} + \lambda t e^{\lambda t} \right) - (\lambda t)^2 = \lambda t.
\]

Fig. 2.3. Derivation of mean and variance of the Poisson distribution.

\[
\prod_{i=0}^{x} U_i > e^{-\lambda t}. 
\]

This will be a sample from the Poisson distribution with parameter \(\lambda\).
2.5 CDF Inversion, Continuous Case: The Exponential Distribution

The exponential distribution is used to model the random waiting time for “events” to occur. Examples of what we mean by events are customers arriving at a queue, the failure of mechanical components, and biological births or deaths. The assumptions about these events are the same as those made for the Poisson distribution. In fact, there is an intimate connection between the two.

The exponential random variable is the waiting time (or distance) for the first Poisson event to occur. Since the probability that a Poisson event will not occur in the interval $[0, t]$ is $e^{-\lambda t}$, as seen by taking $x = 0$ events in (2.9), we have

$$\Pr(\text{an event occurs in } [0, t]) = 1 - \Pr(\text{no event occurs in } [0, t]) = 1 - e^{-\lambda t}. \quad (2.10)$$

This is the cdf of the exponential distribution,

$$\text{cdf}(t) = F(t) = 1 - e^{-\lambda t}.$$

The exponential distribution has a single parameter $\lambda$, the event rate; we denote the distribution by $E(\lambda)$. By its nature as the time at which an event occurs, the exponential is a continuous distribution. Therefore the exponential density is the derivative of its cdf,

$$\text{pdf}(t) = F'(t) = \lambda e^{-\lambda t}.$$

Likewise, probabilities are given by integrals. Let $W$ be the random waiting time before an exponential event occurs. Then the probability that $W$ occurs in some subset $S$ of the real numbers is

$$\Pr(W \in S) = \int_S \lambda e^{-\lambda t} \, dt.$$

Properties of the Exponential Distribution

The exponential distribution is “memoryless” (also called Markovian). For example, if an event has not yet occurred by time $a$, then the probability that it occurs in the interval $a$ to $a + b$ is the same as that of its occurring in $0$ to $b$,

$$\Pr(W < a + b \mid W \geq a) = \Pr(W < b).$$
This is so because in terms of the cdf, the left-hand side is exactly
\[
\frac{F(a + b) - F(a)}{1 - F(a)} = \frac{(1 - e^{-\lambda(a+b)}) - (1 - e^{-\lambda a})}{1 - (1 - e^{-\lambda a})} = \frac{e^{-\lambda a} - e^{-\lambda a} e^{-\lambda b}}{e^{-\lambda a}} = 1 - e^{-\lambda b}.
\]

The mean of the exponential is \( \mu = 1/\lambda \), because using integration by parts with \( u = t \) and \( dv = \lambda e^{-\lambda t} \), we have
\[
\mu = \mathbb{E}(W) = \int_0^\infty t\lambda e^{-\lambda t} dt = -te^{-\lambda t}\bigg|_0^\infty - \int_0^\infty -e^{-\lambda t} dt
\]
\[
= 0 + \frac{e^{-\lambda t}}{-\lambda}\bigg|_0^\infty = \frac{1}{\lambda}. \tag{2.11}
\]

A similar calculation shows that the variance is
\[
\text{var} = 1/\lambda^2; \tag{2.12}
\]
see Figure 2.4.

Calculating the variance of the exponential distribution is done using integration by parts twice,
\[
\text{var} = \mathbb{E}((W - \mu)^2) = \int_0^\infty (t - \frac{1}{\lambda})^2 \lambda e^{-\lambda t} dt
\]
\[
= -(t - \frac{1}{\lambda})^2 e^{-\lambda t}\bigg|_0^\infty + \int_0^\infty 2(t - \frac{1}{\lambda})e^{-\lambda t} dt
\]
\[
= \frac{1}{\lambda^2} + 2 \left[ -\frac{1}{\lambda}(t - \frac{1}{\lambda})e^{-\lambda t}\bigg|_0^\infty + \frac{1}{\lambda} \int_0^\infty e^{-\lambda t} dt \right]
\]
\[
= \frac{1}{\lambda^2} + 2 \left[ -\frac{1}{\lambda^2} + \frac{-1}{\lambda^2} e^{-\lambda t}\bigg|_0^\infty \right] = \frac{1}{\lambda^2}.
\]

\textbf{Fig. 2.4.} Derivation of the variance of the exponential distribution.
2.5.1 Inverting the CDF—The Canonical Method for the Exponential

The method for obtaining a sample from $E(\lambda)$ is simple. Let $U \sim U(0, 1)$ and solve $U = 1 - e^{-\lambda W}$ for $W$. This will give

$$W = \frac{-1}{\lambda} \ln(1 - U),$$

and $W$ is the desired sample. Note that most computer random number generators return values strictly less than 1, in which case the argument of the logarithm will not be 0. This is a general-purpose method, and its explanation is given by the following theorem.

**Theorem 2.1.** (Inverting the cumulative distribution function, continuous case.) Let $F(x)$ be the cdf for the random variable $X$, suppose that $F$ is continuous, and let $U = F(X)$. Then $U$ is uniformly distributed on $[0, 1)$.

**Proof.** We need to show that $\Pr(U \leq u) = u$ for $0 \leq u < 1$. Since $F$ is continuous and increasing, $F^{-1}$ exists. If $x = F^{-1}(u)$ then

$$\Pr(U \leq u) = \Pr(F(X) \leq u) = \Pr(X \leq F^{-1}(u)) = \Pr(X \leq x) = F(x) = u,$$

as claimed. So if $U$ is a sample from $U(0, 1)$, then $X = F^{-1}(U)$ is a sample from cdf $F$.

Inverting the cdf also works for a distribution that is a mixture of discrete and continuous components.

**Inverting the cumulative distribution, general case.**

Sample $U \sim U(0, 1)$; return

$$X = \inf F^{-1}([U, 1]) = \inf \{\xi : F(\xi) \geq U\}.$$

For a proof of this, refer to Figure 2.5.

The main drawback to the method of inverting the cdf is that this usually requires an explicit formula for the cdf, which is often not available. However, in the cases in which it is available it is a very powerful technique.
If $F$ is continuous, this becomes the same as in Theorem 2.1. Where $F$ has jumps, note that $F$ has to be right continuous, since
\[
\lim_{\epsilon \to 0^+} \Pr(X \leq x + \epsilon) = \Pr(X \leq x).
\]
Now assume for the moment the statement $X \leq x$ if and only if $U \leq F(x)$. Then we have
\[
\Pr(X \leq x) = \Pr(U \leq F(x)) = F(x),
\]
which was to be shown. As to the assumption, if $U \leq F(x)$ then $x \in \{\xi : F(\xi) \geq U\}$, so $x \geq X$, which is the infimum of this set. On the other hand, if $\bar{X} \leq x$, then by monotonicity, $F(X) \leq F(x)$, so $U \leq F(x)$.

**Fig. 2.5.** Proof of the *inverting the cdf* technique in the general case.

**Sampling from the Poisson Distribution**

Since the Poisson random variable is the number of exponential events that occur in the interval $[0, t)$, we may simulate the Poisson distribution by sampling exponentials until time $t$ is exceeded. We give here three methods to do this.

**Method 1.** Let $W_i$ be $E(\lambda)$ and let $K$ be the greatest integer such that $\sum_{i=1}^{K} W_i < t$.

**Method 2.** Let $Y_i$ be $E(1)$ and let $K$ be the greatest integer such that $\sum_{i=1}^{K} Y_i < \lambda t$. This works because if $Y$ is $E(1)$, then $X = (1/\lambda)Y$ is $E(\lambda)$ (see below) and $\sum_{i=1}^{K} Y_i < \lambda t$ if and only if $\sum_{i=1}^{K} X_i < t$. This method is an improvement, since one does not have to divide by $\lambda$ on each sample.

**Method 3.** (From Method 2 taking logs.) Let $U_0 = 1$ and $U_i \sim U(0, 1)$ for $i = 1, 2, \ldots$. Let $K$ be the greatest integer such that
\[
\prod_{i=0}^{K} U_i > e^{-\lambda t}.
\]
This works because
\[
-\sum_{i=0}^{K} Y_i > -\lambda t \quad \text{if and only if} \quad \prod_{i=0}^{K} U_i > e^{-\lambda t},
\]
and $Y_i$ is $E(1)$ if and only if $e^{-Y_i}$ is $U(0, 1)$. 

Multiple Independent Waiting Times

Suppose process $A$ has waiting time $W_A$ given by an exponential distribution with parameter $\lambda_A$, and process $B$ has waiting time $W_B$ given by an exponential distribution with parameter $\lambda_B$. The question is, What is the distribution for the process consisting of the first of $A$ or $B$ to occur? We assume that processes $A$ and $B$ are independent.

One can see that

$$\Pr(A \text{ or } B \text{ occurs in } [t, t + \Delta t] | \text{not before}) \approx \lambda_A \Delta t + \lambda_B \Delta t,$$

with the approximation becoming exact as $\Delta t \to 0$. Hence the event time for the combined process is exponential with parameter $\lambda_A + \lambda_B$.

Alternatively, we can argue from the cdfs thus:

$$\Pr(W_A \text{ or } W_B < t) = \Pr(W_A < t) + \Pr(W_B < t) - \Pr(W_A \text{ and } W_B < t)$$

$$= (1 - e^{-\lambda_A t}) + (1 - e^{-\lambda_B t}) - (1 - e^{-\lambda_A t})(1 - e^{-\lambda_B t})$$

$$= 1 - e^{-(\lambda_A + \lambda_B) t},$$

giving the same result.

2.5.2 Discrete Event Simulation

There are many situations in business and industry in which it is necessary to analyze processes that involve stochastic waiting times. This is the field of queuing theory. If one wishes to have an exact analysis of such a process, it is necessary for the waiting times to be exponentially distributed. This is so because the exponential is the only distribution that is “memoryless.” That is, the future depends only on the present and not on how the present was arrived at.

Even so, anything beyond the simplest setup is too hard to analyze directly, so one turns to Monte Carlo simulation; it is known as discrete event simulation in this field. Among its many benefits, simulation allows any waiting time distribution to be used as easily as any other.

Due to its importance, there are several mature software solutions dedicated to providing answers for the most intricate of queuing situations. Much has been written about the area including the important mathematical problem of analyzing simulation results. In this section we give a necessarily brief description of one of the main approaches to discrete event simulation, that of event scheduling.
Event Scheduling Simulation

A software simulation is managed by a “simulation executive” that is responsible for stepping time through the simulation and invoking the other modules as necessary. These include modules for processing the entities of the simulation, implementing their interdependencies, and data display and logging modules.

A timetable of events is maintained, always in chronological order, and the simulation proceeds from one event to the next event in the schedule. Each event has a tag describing when it occurs in absolute time and its nature. When an event is handled, it may generate new events each of which is scheduled by sampling from the appropriate waiting time distribution and merging it into the master schedule at the appointed time.

Consider the example of simulating cell tissue growth, say starting with one cell at time 0 and carrying the simulation through for 8 hours. Assume that cell maturation times are exponentially distributed with parameter $\lambda = 0.5$ per hour. The master schedule needs to contain event times only, since all the events are the same, namely binary fission of identical cells. A more ambitious simulation could invoke crowding and other effects between the cells as desired. Drawing a sample from $E(0.5)$, we might get $W = 3$ hours, say. So our original cell divides at time $t = 3$ and our master schedule is $S[1] = 3$.

Now move time forward to $t = 3$ and handle the cell division of the first cell. Sample $E(0.5)$ again to get $W = 4$, say, for the first daughter cell, so now $S[2] = 7$ (the present time is $t = 3$ and the cell division time is 4, so the event is at absolute time 7). And again for the second daughter cell, get $W = 1$, say. Now we must shift the $S[2] = 7$ down to keep the schedule in time-sorted order, so $S[2] = 4$ and $S[3] = 7$.

The next event on the schedule is at time $t = 4$, the division of a first-generation daughter cell. Sample from $E(0.5)$ to get $W = 2$, say, and again to get $W = 5$. For the $W = 2$ event, its absolute time will be $4 + 2 = 6$, so the master schedule becomes $S[3] = 6$ and $S[4] = 7$.

The $W = 5$ event occurs at absolute time $t = 9$, which is beyond our study, so we need not include it in the master schedule. (But it would have to be accounted for if cell interactions were being monitored.)

Continuing in this fashion, we finally reach the point where all subsequent events occur beyond absolute time $t = 8$ and the simulation is over. At this point the statistics of the run are tabulated and presented.
2.5.3 Transforming Random Variables: The Cauchy Distribution

In the foregoing it was stated that if \( Y \) is \( E(1) \), then \( X = (1/\lambda)Y \) is \( E(\lambda) \). Transforming or mapping random variables like this occurs frequently and is quite useful. In complete generality, a transformation of \( Y \) to \( X \) is given by \( X = g(Y) \) for some one-to-one function \( g \). To find the distribution of \( X \) (or to verify that it has some stated distribution) one can examine either the pdf or the cdf of \( X \).

In our specific example we know that cdf \( Y(y) = P(Y \leq y) = 1 - e^{-y} \), since \( Y \) is \( E(1) \). Therefore we have the following chain of equalities:

\[
cdf_X(x) = Pr(X \leq x) = Pr\left(\frac{1}{\lambda}Y \leq x\right) = Pr(Y \leq \lambda x) = 1 - e^{-(\lambda x)}.
\]

Thus the cdf of \( X \) is \( 1 - e^{-\lambda x} \), so \( X \) is \( E(\lambda) \).

Alternatively, the pdf can be matched. In order to carry this out we have to recall the change of variables technique in integration. Let \( t = g(\tau) \) and \( A = [a, b] \) be an interval. Put \( B = g^{-1}(A) \) (recall \( g \) is one-to-one). Then

\[
\int_A f(t) \, dt = \int_B f(g(\tau))g'(\tau) \, d\tau.
\] (2.13)

The application to densities goes like this. With \( X = g(Y) \),

\[
Pr(X \in A) = Pr(g(Y) \in A) = Pr(Y \in g^{-1}(A)) = \int_{g^{-1}(A)} f_Y(\tau) \, d\tau,
\]

where \( f_Y \) is the pdf for \( Y \). But by a change of variable,

\[
Pr(X \in A) = \int_A f_X(t) \, dt = \int_{g^{-1}(A)} f_X(g(\tau))g'(\tau) \, d\tau,
\]

where \( f_X \) is the pdf of \( X \). Equating the two end members yields

\[
f_Y(\tau) = f_X(g(\tau))g'(\tau). \tag{2.14}
\]

For our exponential mapping, \( f_Y(\tau) = e^{-\tau}, g(\tau) = \tau/\lambda \), and, we anticipate, \( f_X(\tau) = \lambda e^{-\lambda \tau} \). Substitution into (2.14) gives

\[
e^{-\tau} = \lambda e^{-\lambda (\tau/\lambda)} \frac{1}{\lambda},
\]
which is an identity, and thus we have verified the pdf of $X$.

Unfortunately, (2.14) is implicit for $f_X$. To get an explicit expression, we have to apply the change of variable to $g^{-1}$ instead of to $g$. Doing so gives

$$f_X(x) = f_Y(g^{-1}(x)) \left(g^{-1}\right)'(x). \quad (2.15)$$

As a further example we work out the Cauchy density $f_X(t)$. Let $Y \sim U(-\pi/2, \pi/2)$ and $X = \tan(Y)$. Then we can compute that

$$f_X(t) = \frac{1}{\pi} \frac{1}{1 + t^2}. \quad (2.16)$$

To see this, we notice that $g(t) = \tan(t)$, so that $g^{-1}(t) = \arctan(t)$ and thus

$$f_X(t) = \frac{1}{\pi} \frac{d}{dt} (\arctan(t)) = \frac{1}{\pi} \frac{1}{1 + t^2},$$

as desired.

The Cauchy density function

![The Cauchy density function](image)

Fig. 2.6. The Cauchy probability density function.

The Cauchy density is symmetric about $t = 0$, and so its mean $\mu_X$ is 0; see Figure 2.6. But the variance of the Cauchy distribution is infinite, since the integral

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{t^2}{1 + t^2} dt$$
is unbounded. Before Gauss’s work, the Cauchy distribution was thought to be a candidate for the normal distribution, i.e., the distribution of the central limit theorem. Although the pdfs of the two distributions have a similar appearance, the tails of the Cauchy distribution are too “fat,” as evidenced by the fact that the normal distribution has finite variance but the Cauchy distribution does not.

2.6 The Central Limit Theorem and the Normal Distribution

In many of the histograms generated by the problems we have worked on up to now, the figure that has emerged is that of a bell-shaped curve. This is not by accident; it is the consequence predicted by the central limit theorem, which we now consider.

The graph of the function \( f(x) = e^{-x^2/2} \) (see Figure 2.7) is symmetric about the origin, everywhere positive, bell-shaped, and tends to 0 as \( x \to \pm \infty \). Further, the integral of \( f \) over the whole real line is \( \sqrt{2\pi} \),\(^1\) and therefore

\[
 f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}
\]

is a probability density function.

Replacing \( x \) by \( x/\sigma \), where \( \sigma \) is a parameter, has the effect of rescaling the \( x \)-axis in terms of \( \sigma \); that is, the rescaled function at \( x = 2\sigma \) has the same value as the original function at \( x = 2 \). As a side effect, to maintain a unit integral, \( f \) must be divided by \( \sigma \) as well.

Finally, replacing \( x \) by \( x - \mu \) has the effect of shifting the function; the shifted function at \( x = \mu \) has the same value as the original at \( x = 0 \).

With these changes, the normal probability density function is

\[
 f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}. \tag{2.17}
\]

Denote this distribution by \( N(\mu, \sigma^2) \). By the transformation principles for random variables (from the last section), it is easy to see that if \( X \) is distributed as \( N(\mu, \sigma^2) \), then \( Y = (X - \mu)/\sigma \) is distributed as \( N(0, 1) \).

The mean and variance of the \( N(\mu, \sigma^2) \) distribution are, appropriately enough, \( \mu \) and \( \sigma^2 \) respectively. The calculation is given in Figure 2.8.

\(^1\) This can be seen by calculating \( \int_{-\infty}^{\infty} e^{-x^2/2} dx \int_{-\infty}^{\infty} e^{-y^2/2} dy \) in polar coordinates.
2.6 The Central Limit Theorem and the Normal Distribution

The $N(0,1)$ density function.

\[ \text{Fig. 2.7. The standard normal density function.} \]

Obtaining the cumulative distribution function for the normal distribution is harder to come by. There is no antiderivative for $e^{-x^2}$ in terms of elementary functions and so there is no simple form for the normal cumulative distribution function. However, approximate values may be computed by the technique of rational approximation, that is, the cdf of $N(0,1)$ is approximated by a rational function. Table 2.2 was computed using the code below, which uses this technique. Amazingly, the rational approximation produces values correct to three decimal places for all values of $x$, as in the table.

Matlab
```matlab
function w = normalCDF(x)
    a1 = 0.31938153; a2 = -0.356563782;
    a3 = 1.781477937; a4 = -1.821255978;
    a5 = 1.330274429; b=0.2316410; c=1/sqrt(2*pi);
    L=abs(x); K=1./(1+b.*L);
    w = 1-c*exp(-L.*L/2).*...
        (a1.*K+a2.*K.^2+a3.*K.^3+a4.*K.^4+a5.*K.^5);
    msign = -(x<0); mfact = 2*msign+1;
    w = mfact.*(w+msign);
```

Theorem 2.2. (Central limit theorem) Let $X_1, X_2, \ldots, X_n$ be independent random samples from a distribution with mean $\mu$ and finite variance $\sigma^2$. Then
Let \( u = (x - \mu)/\sigma \); then \( x = \sigma u + \mu \) and \( du = dx/\sigma \) and we have

\[
\mathbb{E}(X) = \int_{-\infty}^{\infty} \frac{x}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, du = \int_{-\infty}^{\infty} \frac{\sigma u + \mu}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du
\]

\[
= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{1}{2}u^2} \, du + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du
\]

\[
= \mu,
\]

since the first integral is 0 by symmetry and the second integral is 1 since \( f \) is a pdf. To calculate the variance we must add integration by parts to the substitution used above,

\[
\mathbb{E}(X^2) = \int_{-\infty}^{\infty} \frac{x^2}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, du
\]

\[
= \int_{-\infty}^{\infty} \frac{\sigma^2 u^2 + 2\sigma \mu u + \mu^2}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du
\]

\[
= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{1}{2}u^2} \, du + \frac{2\sigma \mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{1}{2}u^2} \, du + \mu^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du.
\]

The second member on the last line is zero by symmetry, and the third integrates to \( \mu^2 \). For the first member, use integration by parts and get

\[
\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{1}{2}u^2} \, du = \frac{\sigma^2}{\sqrt{2\pi}} \left[ u e^{-\frac{1}{2}u^2} \right]_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} \, du
\]

\[
= \frac{\sigma^2}{\sqrt{2\pi}} (0) + \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du
\]

\[
= \sigma^2.
\]

Therefore the variance is \( \sigma^2 \), since

\[
E \left( (X - \mu)^2 \right) = \mathbb{E}(X^2) - \mu^2 = \sigma^2.
\]

**Fig. 2.8.** Mean and variance calculation for the normal distribution.
### 2.6 The Central Limit Theorem and the Normal Distribution

#### Table 2.2. Computed approximate normal cdf.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F(x)$</td>
<td>0.500</td>
<td>0.540</td>
<td>0.579</td>
<td>0.618</td>
<td>0.655</td>
<td>0.691</td>
<td>0.726</td>
<td>0.758</td>
<td>0.788</td>
<td>0.816</td>
</tr>
<tr>
<td>$x$</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.3</td>
<td>1.4</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
<td>1.8</td>
<td>1.9</td>
</tr>
<tr>
<td>$F(x)$</td>
<td>0.841</td>
<td>0.864</td>
<td>0.885</td>
<td>0.903</td>
<td>0.919</td>
<td>0.933</td>
<td>0.945</td>
<td>0.955</td>
<td>0.964</td>
<td>0.971</td>
</tr>
<tr>
<td>$x$</td>
<td>2.0</td>
<td>2.1</td>
<td>2.2</td>
<td>2.3</td>
<td>2.4</td>
<td>2.5</td>
<td>2.6</td>
<td>2.7</td>
<td>2.8</td>
<td>2.9</td>
</tr>
<tr>
<td>$F(x)$</td>
<td>0.977</td>
<td>0.982</td>
<td>0.986</td>
<td>0.989</td>
<td>0.992</td>
<td>0.994</td>
<td>0.995</td>
<td>0.997</td>
<td>0.997</td>
<td>0.998</td>
</tr>
</tbody>
</table>

\[ Y = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \]

has a limiting distribution as $n \to \infty$ and it is $N(0,1)$, normal with mean 0 and variance 1.

Exercise 29 gives empirical evidence for the central limit theorem.

#### 2.6.1 Sampling from the Normal Distribution

We discuss three methods for sampling from the normal distribution. For a comparison of these methods, see Problem 26 at the end of this chapter.

Our first method takes advantage of the central limit theorem and is only crudely approximate. Its advantage is that it is quick and easy to program.

**Central Limit Algorithm**

Sample $n$ copies of $U(0,1)$ random variables, $U_1, U_2, \ldots, U_n$, and return

\[ X = \frac{\sum_{i=1}^{n} U_i - n\mu}{\sqrt{n/12}} \]

By the central limit theorem, $X$ is approximately normal with parameters $\mu = 0$ and $\sigma = 1$, that is, $N(0,1)$. Recall from equations (1.14) and (1.15) that the mean and variance of a $U(0,1)$ random variable are $1/2$ and $1/12$ respectively. Often $n$ is taken to be 12 to avoid the square root and division.

**Box–Muller Algorithm**

This method gives an exact sample. Sample random variables $U_1, U_2 \sim U(0,1)$ and put
\[X_1 = \cos(2\pi U_1)\sqrt{-2\ln U_2},\]
\[X_2 = \sin(2\pi U_1)\sqrt{-2\ln U_2}.\]  (2.18)

Then \(X_1\) and \(X_2\) are independent normally distributed random variables with mean 0 and variance 1. The proof of this is given in Figure 2.10.

Although it might seem that knowing \(X_1\) imparts some information about \(X_2\), this is not so; \(X_2\) can still be any real value at all. An example is shown in Figure 2.9. The graph labeled “\(X_1 = 0.7\)” graphs the points \((U_1, U_2)\) in the plane satisfying the first of equations (2.18) for a given value of \(X_1\), in this case 0.7. The graph labeled “\(X_2 = 2.0\)” shows the points satisfying the second of equations (2.18) for the given value of \(X_2\). As shown, there is a unique pair \((U_1, U_2)\) giving these choices.

\[
\begin{align*}
X_1 &= 0.7 \\
X_2 &= 2.0 \\
X_1 &= 0.7
\end{align*}
\]

\[
\begin{align*}
0 & \quad 0.2 & \quad 0.4 & \quad 0.6 & \quad 0.8 & \quad 1 \\
U_1 & \quad & & & & \\
U_2 & \quad 0 & \quad 0.1 & \quad 0.2 & \quad 0.3 & \quad 0.4 & \quad 0.5 & \quad 0.6 & \quad 0.7 & \quad 0.8 & \quad 1
\end{align*}
\]

Fig. 2.9. Solving for \((U_1, U_2)\) for a given \((X_1, X_2)\).

**Marsaglia–Bray Algorithm**

This method also gives two exact samples, similar to the Box–Muller algorithm.

1. Sample random variables \(U_1, U_2 \sim U(0, 1)\) and put \(V_1 = 2U_1 - 1\), \(V_2 = 2U_2 - 1\), and \(S = V_1^2 + V_2^2\).
2. if $S \geq 1$ go to 1, else return

$$X_1 = V_1 \left( \frac{-2}{S} \ln S \right)^{1/2},$$

$$X_2 = V_2 \left( \frac{-2}{S} \ln S \right)^{1/2}.$$ 

Then $X_1$ and $X_2$ are independent normally distributed random variables with mean 0 and variance 1.

The Marsaglia–Bray algorithm was derived as an improvement in speed over the Box–Muller algorithm, since it uses only one transcendental function evaluation, while the Box–Muller uses three. This is offset somewhat by the fact that a fraction of the time Step 1 is rejected and must be repeated. (What is this fraction?) Today the speed of the computer and the efficiency of transcendental function evaluation has made these issues moot in most cases.

The proof that the Marsaglia–Bray algorithm gives exact samples from $N(0, 1)$ is similar to that of the Box–Muller algorithm and we omit it.

### 2.6.2 Approximate Sampling via the Central Limit Theorem

Suppose we want a sample of the outcome of 1000 Bernoulli trials with probability of success $p = 0.995$. This might correspond to the number of acceptable light bulbs to come out of a manufacturing process in one hour. Of course we could simulate the 1000 trials. Or we could instead exploit the CLT as follows. Let $R_{1000}$ be the random variable for the number of successes, and let $\mu$ and $\sigma^2$ refer to the Bernoulli trial random variable; $\mu = p$ and $\sigma^2 = pq$. Then $N$ defined by

$$N = \frac{R_{1000} - 1000\mu}{\sqrt{1000\sigma^2}} = \frac{R_{1000} - 995}{2.2305}$$

is approximately $N(0, 1)$. Now let $n$ be a sample from $N(0, 1)$, say $n = 0.68$. Then

$$r = 2.2305n + 995 = 996.52 \rightarrow 997,$$

rounded to an integer, is the corresponding sample for $R_{1000}$. 
Proof for the Box–Muller algorithm. Let \( f_{X_1, X_2}(x_1, x_2) \) be the joint density of \( X_1 \) and \( X_2 \) and let \( f_{U_1, U_2}(u_1, u_2) \) be the joint density of \( U_1 \) and \( U_2 \). Then

\[
\begin{align*}
  f_{U_1, U_2}(u_1, u_2) &= f_{U_1}(u_1) f_{U_2}(u_2) = 1, \quad \text{if } 0 \leq u_1, u_2 < 1, \\
  \text{and } 0 \text{ otherwise.}
\end{align*}
\]

and \( \tau \) be the change of variable

\[
\begin{bmatrix}
  x_1 \\
  x_2
\end{bmatrix} = \tau(u_1, u_2) = \left( \begin{array}{c}
\cos(2\pi u_1) \sqrt{-2 \ln u_2} \\
\sin(2\pi u_1) \sqrt{-2 \ln u_2}
\end{array} \right).
\]

Squaring \( x_1 \) and \( x_2 \) and adding gives

\[
x_1^2 + x_2^2 = -2(\ln u_2) \cos^2(2\pi u_1) - 2(\ln u_2) \sin^2(2\pi u_1) = -2 \ln u_2.
\]

So

\[
u_2 = e^{-\frac{1}{2}(x_1^2 + x_2^2)}.
\]

This shows that \( 0 < u_2 < 1 \). The variable \( u_1 \) may be similarly restricted due to the periodicity of sine and cosine. Similar to the 1-variable change of variable formulas derived previously, the relationship between the two densities may be written either as

\[
f_{X_1, X_2}(\tau(u_1, u_2)) \det \tau' = f_{U_1, U_2}(u_1, u_2)
\]

or equivalently as

\[
f_{X_1, X_2}(x_1, x_2) = f_{U_1, U_2}(\tau^{-1}(x_1, x_2)) \det(\tau^{-1})'.
\]

Since \( f_{U_1, U_2} \) is identically 1, and remembering that \( \det(\tau^{-1})' = 1/\det \tau' \), the latter gives

\[
f_{X_1, X_2}(x_1, x_2) = \det(\tau^{-1})' = 1/\det \tau'.
\]

Now

\[
\det \tau' = \det \left( \begin{array}{cc}
\frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} \\
\frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2}
\end{array} \right)
\]

\[
= \det \left( \begin{array}{cc}
-2\pi \sin(2\pi u_1) \sqrt{-2 \ln u_2} & -\frac{1}{\sqrt{-2 \ln u_2}} \cos(2\pi u_1) \\
2\pi \cos(2\pi u_1) \sqrt{-2 \ln u_2} & \frac{1}{\sqrt{-2 \ln u_2}} \sin(2\pi u_1)
\end{array} \right)
\]

\[
= \frac{2\pi}{u_2} \sin^2(2\pi u_1) + \frac{2\pi}{u_2} \cos^2(2\pi u_1) = \frac{2\pi}{u_2}
\]

From equation (*), \( u_2 = e^{-\frac{1}{2}(x_1^2 + x_2^2)} \). Therefore

\[
\det(\tau^{-1})' = \frac{u_2}{2\pi} = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)}
\]

and

\[
f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2}.
\]

QED.

Fig. 2.10. Proof for the Box–Muller algorithm.
2.6.3 Error Estimates for Monte Carlo Simulations

Suppose we are trying to use Monte Carlo to estimate some value, call it $\theta$. It could be the waiting time of a discrete event simulation or the neutron flux through a thickness of shielding or any scalar value that results from an instance of a simulation.

Let $X_1, X_2, \ldots, X_n$ be $n$ estimates of $\theta$ as derived from the outcome of the simulation. If the $X_i$ are independent and identically distributed with mean $\theta$, then by the central limit theorem their sample average $\bar{X}$ is approximately normally distributed with mean equal to $\theta$ and variance equal to $\sigma^2_X/n$, where $\sigma^2_X$ is the (unknown) variance of the $X_i$. In this case

$$Y = \frac{\bar{X} - \theta}{\sqrt{\sigma^2_X/n}}$$

is approximately $N(0,1)$ distributed. From a $N(0,1)$ table we notice that among other things, with probability 0.954 a normal sample lies within two standard deviations of the mean; hence

$$\Pr \left( -2 < \frac{\bar{X} - \theta}{\sqrt{\sigma^2_X/n}} < 2 \right) = 0.954.$$ 

In other words, with probability 0.954, $\theta$ lies in the interval

$$\bar{X} - 2\sqrt{\sigma^2_X/n} < \theta < \bar{X} + 2\sqrt{\sigma^2_X/n}.$$ 

Now, given a value for $\sigma^2_X$, we may calculate probabilistic error bounds for $\theta$, or confidence intervals as they are called.

In practice there are three problems with this program. First, usually $\sigma^2_X$ must itself be estimated from the data. In that case $Y$ above will not be normal. Second, the $X_i$ may not be identically distributed; the simulation may suffer start-up effects, for example. And third, the $X_i$ may be correlated. Both of these issues are, in fact, a common difficulty in queuing simulations.

An attempt to deal with the second and third issues is by batching. Divide the $n$ trials into $m$ batches each of size $J$:

$$X_1 \ldots X_J \mid X_{J+1} \ldots X_{2J} \mid \ldots \mid X_{(m-1)J+1} \ldots X_{mJ}.$$ 

Thus there are $m = n/J$ batches. The batch random variables
\[ B_i = \frac{1}{J} \sum_{j=(i-1)J+1}^{iJ} X_j, \quad i = 1, \ldots, m, \]
tend to be independent and identically distributed. Now we may apply the development above to the \( B_i \) in place of the \( X_i \). Thus

\[ \hat{\theta} = \bar{B} = \frac{1}{m} \sum_{i=1}^{m} B_i \]
is an estimator \( \hat{\theta} \) for \( \theta \). And the random variable

\[ Y = \frac{\hat{\theta} - \theta}{\sqrt{\sigma_B^2/m}} \]
is approximately \( N(0, 1) \).

If we knew the variance \( \sigma_B^2 \) of the batch random variables, then we could use the normal distribution itself to make error bounds as was done above. However, \( \sigma_B^2 \) is generally not known and must itself be estimated from the \( B_i \) data. The sample variance for \( \sigma_B^2 \) is given by

\[ s_B^2 = \frac{1}{m-1} \sum_{i=1}^{m} (B_i - \hat{\theta})^2 \quad (2.19) \]
and is itself a random variable (known to be gamma distributed; see Section 2.8).

Thus in place of \( Y \) we have

\[ t = \frac{\hat{\theta} - \theta}{(s_B/\sqrt{m})}, \quad (2.20) \]
the quotient of a normal random variable by a gamma random variable; see [HT01]. Such a combination is a Student-\( t \) random variable. The Student-\( t \) is well known and has one parameter, its degrees-of-freedom or DOF. Tables giving ordinates of the Student-\( t \) are included in the appendix of most statistics books. We provide an abbreviated one as well, Table 2.4. So we must use the \( t \)-statistic to obtain the confidence intervals we seek.

For example, given \( \alpha \), \( t_\alpha \) is defined by

\[ \Pr(-t_\alpha < t < t_\alpha) = \alpha. \]
These values $t_\alpha$ can be looked up in or derived from a $t$-table. If the table gives cumulative values instead of two-sided values, as does the table below, a conversion is required. Since the $t$ distribution is symmetric, $\Pr(-t_\alpha < t < t_\alpha) = \alpha$ is equivalent to

$$\Pr(t < t_\alpha) = \alpha + \frac{1 - \alpha}{2} = \frac{1 + \alpha}{2}.$$ 

Thus if one wants, say, a 95% confidence interval for $t$, use the cumulative table with $\beta = (1 + 0.95)/2 = 0.975$.

The Student-$t$ distribution table gives $t_\beta$ vs DOF, where $t_\beta$ is defined by $\Pr(t < t_\beta) = \beta$.

<table>
<thead>
<tr>
<th>DOF</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>60</th>
<th>120</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_\beta$</td>
<td>6.31</td>
<td>2.92</td>
<td>2.35</td>
<td>2.13</td>
<td>2.02</td>
<td>1.94</td>
<td>1.86</td>
<td>1.81</td>
<td>1.75</td>
<td>1.70</td>
<td>1.68</td>
<td>1.67</td>
<td>1.66</td>
<td>1.65</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.3.** Cumulative $t$ values for $\beta = 0.95$.

<table>
<thead>
<tr>
<th>DOF</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>60</th>
<th>120</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_\beta$</td>
<td>12.71</td>
<td>4.30</td>
<td>3.18</td>
<td>2.78</td>
<td>2.57</td>
<td>2.45</td>
<td>2.31</td>
<td>2.23</td>
<td>2.13</td>
<td>2.09</td>
<td>2.04</td>
<td>2.02</td>
<td>2.00</td>
<td>1.98</td>
<td>1.96</td>
</tr>
</tbody>
</table>

**Table 2.4.** Cumulative $t$ values for $\beta = 0.975$.

A problem that arises with batching is deciding how to divide up the $n$ samples between the number of batch replications $m$ versus the number of repetitions per batch $J$. We want $m$ as large as possible, since a larger $m$ reduces the size of the confidence interval. This argues for a small value of $J$. But $J$ must be large enough that the batch random variables tend to be independent and uncorrelated.

### 2.7 Rejection Sampling Example: The Beta Distribution

The Marsaglia–Bray algorithm for the normal distribution introduces a new idea into sampling, namely a test imposed on candidate values in order to proceed. This idea can be expanded upon to form the basis of a powerful and completely general sampling method, called rejection sampling.

Consider the problem of sampling a Bernoulli trial with probability $p = 1/3$ using coin flips. Here is one solution: flip the coin twice; if the
outcome is HH, return “success,” if HT or TH, return “failure,” and if TT, reject the trial and go again. This gives the right probabilities because

\[
\Pr(\text{return is success}) = \frac{\Pr(\text{return is success | there is a return})}{\Pr(\text{there is a return})} = \frac{1/4}{3/4} = \frac{1}{3},
\]

By adding the acceptance/rejection discipline, only the relative values of the target probabilities matter. That is, they could have been 1/7, 1/7, 1/7 and “success” will still be returned 1/3 of the time.

Consider a second example. Using three coin tosses, we want to select “1” with probability 2/5, and “2,” “3,” and “4” each with probability 1/5. Since tossing a coin twice gives probabilities of 1/4 and this is less than 2/5, we divide the target probabilities by \( M = \frac{8}{5} \) (to bring 2/5 down to 1/4), recalling that only their relative values matter. The modified target probabilities are then 1/4, and 1/8, 1/8, 1/8. To find the acceptance discipline, reason as follows. If the result HH occurs, we accept this all the time and return “1” because \( \frac{1}{4} \cdot \frac{1}{4} = 1 \). If the result HT occurs, we accept this one-half of the time and map its return to “2”; this gives the correct probability because \( \frac{1}{8} \cdot \frac{1}{4} = \frac{1}{2} \). Similarly we accept TH and TT one-half of the time and map these outcomes to “3” and “4” respectively. If a value is rejected, we start over and toss the two coins again.

The probability of accepting a value is this: all the time for HH, and \( 1/2 \) of the time for the others; hence

\[
\frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{5}{8},
\]

or 62.5% of the time; so 37.5% of the time we must start again. When a value is returned, the probability that it is a “1” is given by

\[
\Pr(\text{returned value is “1”}) = \frac{\Pr(“1” \text{ is selected and returned})}{\Pr(\text{some value is returned})} = \frac{1/4 \cdot 1}{5/8} = \frac{2}{5}.
\]

In the same manner, for “2” we have
Pr(returned value is "2") = \frac{Pr("2" is selected and returned)}{Pr(some value is returned)} = \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{1}{8}} = \frac{1}{5},

and the other target probabilities are likewise seen to have the proper frequencies.

To actually carry out the method, toss a coin twice; if the outcome is HH, then return "1"; if the outcome is HT or TH or TT, then we must do a side experiment to decide whether to accept the outcome and return its mapped value or reject the outcome and start again. Since the side experiment is a 50-50 decision, it could itself be a coin toss; say an H means accept and a T means reject. For example, if the third toss is H then return "2" in the HT case, "3" in the TH case, and "4" in the TT case. If it is T, then start all over.

Figure 2.11 illustrates the situation in the example. For each possible outcome from the first two coin tosses, we must scale the resulting probability to obtain the correct value.

![Diagram](image)

**Fig. 2.11.** Target \( f \) and proposal \( Mg \) densities for the example in the Text.

The density actually sampled from, the two coin tosses in the above examples, is called the proposal density \( g \). It can be any convenient density that is nonzero where the target density \( f \) is nonzero. Let \( h(\cdot) \)
refer to the acceptance discipline. The *generalized rejection sampling method* consists of two simple steps.

**Rejection Sampling**

1. Let $Y$ be a sample from pdf $g$.
2. With probability $h(Y)$ return $X = Y$; otherwise go to (1).

As in the example, step 2 is a Bernoulli trial with success probability $h(Y)$ and is carried out through a side experiment.

Referring again to Figure 2.11, we see that $M = 8/5$ and we must scale the proposal $g$ (in the example, a uniform distribution given by two coin tosses) by $8/5$ in order to have $f \leq Mg$. Then for each outcome drawn from $g$, the proportion of $f$ that lies under this part of $g$ gives the acceptance probability. For the first outcome of “1” we see that $f = 2g$, so this one is always accepted. For the rest, $f = (1/2)2g$ (the shaded part represents only $1/2$ of the vertical bar at these positions), and so we accept only with probability $1/2$.

Rejection sampling is a completely general method for sampling from both discrete and continuous distributions. To see that it works in the continuous case, calculate the infinitesimal probability that the returned value $X$ equals $x$ as follows:

$$\Pr(X = x) \, dx = \frac{\Pr(Y = x \text{ and } x \text{ is returned})}{\Pr(\text{some value is returned})} = \frac{g(x) \, dx \, h(x)}{\int_{-\infty}^{\infty} g(z) h(z) \, dz}.$$  

Hence the probability density function of returned values is

$$pdf_X(x) = \frac{g(x) h(x)}{\int_{-\infty}^{\infty} g(z) h(z) \, dz}.$$  

To construct a sampling algorithm, let $f$ be the target pdf, let $g$ be the proposal pdf, and let the constant $M$ be such that $f(x) \leq Mg(x)$ for all $x$. The acceptance/rejection discipline is

$$h(x) = \frac{f(x)}{(Mg(x))}. \quad (2.21)$$

With this, the pdf of accepted values is

$$\frac{g(x) h(x)}{\int g h} = \frac{f(x)/M}{\int f(z)/M \, dz} = f(x).$$
As above, the probability of an acceptance is

\[ \Pr(Y \text{ is accepted}) = \int_{-\infty}^{\infty} g(z)h(z) \, dz = 1/M. \]

This is the efficiency of the generator, so keep \( M \) as small as possible. Figure 2.12 illustrates the situation for continuous \( f \) and \( g \).

Notice that it is not a problem if \( g(x) = 0 \) for some \( x \), since we only use (2.21) to compute \( h(x) \) for values of \( x \) that are sampled from a distribution with pdf \( g \). The probability of obtaining an \( x \) such that \( g(x) = 0 \) is zero, and thus the situation will never occur.

### 2.7.1 Sampling from the Beta Distribution

The beta distribution, \( \text{Be}(\alpha, \beta) \), is defined as the density given by

\[ f(x) = cx^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1, \]

where the constant \( c \) is chosen to make \( \int_0^1 f(x) \, dx = 1 \). The parameters \( \alpha \) and \( \beta \) can be any positive real values. If either \( \alpha \) or \( \beta \) is less than 1, the density is unbounded but the integral is nonetheless finite. The mean of the beta distribution is

\[ \mu = \frac{\alpha}{\alpha + \beta} \]

and the variance is

\[ \text{var} = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}. \]
Suppose we want to draw samples from $\text{Be}(2,3)$ whose density is 

$$f(x) = 12x(1-x)^2, \quad 0 \leq x \leq 1;$$

see Figure 2.13. We will use the uniform $U(0,1)$ as the proposal; thus $g(x) = 1, \quad 0 \leq x \leq 1$. Since $f(x) \leq 16/9$ (from the figure), we may take $M = 16/9$. Then the acceptance discipline is

$$h(x) = \frac{f(x)}{Mg} = \frac{27}{4}x(1-x)^2.$$ \hspace{1cm} (2.23)

The rejection sampling algorithm becomes (1) sample $U \sim U(0,1)$, and (2) with probability $h(U)$ return $X = U$; otherwise go to (1).

A run might go as follows, $U \sim U(0,1) = 0.8213$, say, so $h(U) = 0.0786$ as given by (2.23). Now, sampling $R \sim U(0,1)$ as the side experiment gives, say, $R = 0.616$. This is not less than or equal to 0.0786 so $U = 0.8213$ is rejected. Retrying, this time $U \sim U(0,1) = 0.384$, say, and $h(U) = 0.4371$. Another call to $U(0,1)$ produces $R = 0.00506$, which is less than or equal to 0.4371, so the side experiment succeeds. This time $U = 0.384$ is accepted and is returned.

![Fig. 2.13. The β-distribution $12x(1-x)^2$.](image)

This example shows that one can view rejection sampling as an example of hit-or-miss. Note that the chosen value $U$ is accepted if $\frac{16}{9}R < f(U)$ and rejected if $\frac{16}{9}R \geq f(U)$. Plot the point $(U, \frac{16}{9}R)$ in Figure 2.13, and we see that $U$ is accepted if it lies below the graph of $f$ and rejected if above (or on) the graph.
2.7.2 Sampling from an Unbounded Beta Distribution

We take up one more example of rejection sampling. This time we seek samples from an unbounded beta distribution. For simplicity take $\alpha = \beta$ with $\alpha < 1$ (so that the density is unbounded at $x = 0, 1$), but the same idea works more generally. The density is

$$f(x) = cx^{\alpha-1}(1-x)^{\alpha-1}, \quad 0 \leq x \leq 1,$$

where $c$ is the normalizing factor. Recall that only relative values of $f$ are needed, so it is not necessary to find $c$. This is an important feature of the rejection sampling method.

As the proposal distribution we introduce a composition dividing the domain into the subintervals $[0, 1/2)$ and $[1/2, 1)$. We will have more to say about compositions in the next section. By symmetry of this $f(x)$ we need only sample $X$ from the interval $[0, 1/2)$ and, with probability $1/2$, return either $X$ or $1 - X$. Therefore as the proposal density we use

$$g(x) = c'x^{\alpha-1}, \quad 0 \leq x \leq 1/2.$$

The cdf is the integral

$$G(x) = \int_0^x c't^{\alpha-1} dt = c'\frac{x^\alpha}{\alpha}.$$

Since $G(1/2) = 1$, we get $c' = \alpha/(1/2)^\alpha$.

To sample from $g$ by cdf inversion entails solving

$$U = \frac{c'}{\alpha}Y^\alpha$$

for $Y$. We get

$$Y = \left(\frac{\alpha}{c'}U\right)^{1/\alpha} = \frac{1}{2}U^{1/\alpha},$$

where we have substituted for $c'$ from above.

Since $\alpha < 1$, the largest value of $(1-x)^{\alpha-1}$ for $0 \leq x \leq 1/2$ occurs when $x = 1/2$. Therefore we may take $M = (1/2)^{\alpha-1}c/c'$, and with this choice we have

$$Mg(x) = (1/2)^{\alpha-1}\frac{c}{c'}c'x^{\alpha-1} = (1/2)^{\alpha-1}cx^{\alpha-1}$$

$$\geq (1-x)^{\alpha-1}cx^{\alpha-1} = f(x), \quad 0 \leq x \leq 1/2.$$

Finally, the acceptance discipline is given by
\[ h(x) = \frac{f(x)}{Mg(x)} = \frac{cx^{\alpha-1}(1-x)^{\alpha-1}}{(1/2)^{\alpha-1}cx^{\alpha-1}} = (2(1-x))^{\alpha-1}. \]

As predicted, \( c \) drops out and its value is not needed.

The algorithm for \( \text{Be}(\alpha, \alpha) \) is then:

1. generate \( U \sim U(0, 1) \) and put \( Y = (1/2)U^{1/\alpha} \);
2. as the side experiment, if \( R \sim U(0, 1) < h(Y) \) then accept \( Y \) and go to step 3; else return to step 1;
3. again sample \( U \sim U(0, 1) \); if \( U < 1/2 \) return \( X = Y \); otherwise return \( X = 1 - Y \).

### 2.8 Composition Example: The Gamma Distribution

A density \( f \) may be constructed from others, say \( f_1, f_2, \ldots, f_n \), by forming the sum

\[ f(x) = p_1 f_1(x) + p_2 f_2(x) + \cdots + p_n f_n(x). \tag{2.24} \]

The weights \( p_1, p_2, \ldots, p_n \) are positive and sum to 1. The mean of a composition density is

\[ \mu = p_1 \mu_1 + p_2 \mu_2 + \cdots + p_n \mu_n, \tag{2.25} \]

and the variance is given by

\[ \text{var} = p_1^2 \text{var}_1 + p_2^2 \text{var}_2 + \cdots + p_n^2 \text{var}_n, \tag{2.26} \]

since we take the constituent distributions \( f_i \) to be independent. In this \( \mu_i \) and \( \text{var}_i \) are the mean and variance of the \( i \)th distribution.

To sample from a composite distribution, pick index \( i \) as in roulette wheel selection with probability \( p_i \), for example via discrete cdf inversion or the alias method. Having chosen \( i \), sample from \( f_i \) by an appropriate method for that density and return the sampled value.

Composition can give rise to a great variety of probability distributions. One class of examples is the sum of piecewise constant functions such as a histogram. In this case the \( f_i \) are just uniform densities and are easily sampled. Any density can be approximated in this way. A better approximation is to use piecewise trapezoidal regions. In this case the \( f_i \) are appropriately scaled linear functions and these too are easily sampled.
Another way in which composition can be exploited for sampling purposes is by tailoring a piecewise envelope or proposal density \( g \) for use in the rejection method. This gives rise to an exact sampling method. The combination of composition and rejection provides a powerful tool for many distributions that cannot be treated in any other way.

We already saw a restricted example of this approach in the last section, restricted in the sense that by symmetry the two components were essentially the same distribution. Here we show how the composition/rejection technique can provide a solution for treating the gamma distribution. This is another important distribution that arises in conjunction with the Poisson process.

### 2.8.1 The Gamma Distribution

In the Poisson process with event rate \( \lambda \) we saw that the exponential distribution is the waiting time until the first event. By way of generalization, the \textit{gamma distribution} with parameter \( \alpha \) is the waiting time \( W \) until the \( \alpha \)th event occurs. Using (2.9) we obtain the gamma cdf as follows:

\[
F(w) = \Pr(W \leq w) = 1 - \Pr(W > w) \\
= 1 - \Pr(\text{fewer than } \alpha \text{ events occur in } [0, w]) \\
= 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}}{k!}.
\]  

(2.27)

Differentiating this with respect to \( w \), we obtain the gamma density function

\[
f(w) = \frac{\lambda}{(\alpha - 1)!} (\lambda w)^{\alpha-1} e^{-\lambda w}.
\]  

(2.28)

Denote this distribution by \( G(\alpha, \lambda) \).

To begin our treatment of the gamma distribution we first simplify it by means of a transformation inspired by our work with the exponential distribution. It is easy to see that if \( W \sim G(\alpha, \lambda) \), then \( Y = \lambda W \) is distributed as \( G(\alpha, 1) \). Therefore we will restrict our attention to \( G(\alpha, 1) \) and use the inverse transformation, \( W = (1/\lambda)Y \), to get back to the more general version if necessary.

One of the most important properties of the gamma distribution is the \textit{reproductive property}:
if \( W_1 \sim G(\alpha_1, \lambda) \) and \( W_2 \sim G(\alpha_2, \lambda) \) are independent, then \( W_1 + W_2 \) is distributed as \( G(\alpha_1 + \alpha_2, \lambda) \). (See Figure 2.14 for a proof in the integer case.)

This property is to be expected. The waiting time for \( \alpha \) events to occur is just the \( \alpha \)-fold sum of the waiting time for the events to occur one by one. By the reproductive property it follows that the mean and variance of the gamma are just the \( \alpha \)-fold sums of the mean and variance of the exponential; hence

\[
\mu = \frac{\alpha}{\lambda}, \quad \text{var} = \frac{\alpha}{\lambda^2}.
\]

As we will see shortly, the reproductive property will also allow us to generate samples from the gamma distribution by sampling exponentials.

The gamma distribution with \( \alpha \) an integer is also known as the *Erlang distribution*. The gamma distribution extends the Erlang to the case that \( \alpha \) is any positive real number. For nonintegral \( \alpha \) the factorial is replaced by the *gamma function*

\[
\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} \, dy, \quad 0 < t.
\]  

(2.29)

For integer arguments \( n = 1, 2, \ldots \),

\[
\Gamma(n) = (n - 1)!. 
\]

For noninteger arguments the value is given by the integral directly (2.29). In particular, it can be shown that

\[
\Gamma(\frac{1}{2}) = \sqrt{\pi}.
\]

### 2.8.2 Sampling from \( G(\alpha, 1) \)

First let \( 0 < \alpha < 1 \). For purposes of rejection, the target density is

\[
f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}.
\]

Since only relative values count, we can ignore the constant \( 1/\Gamma(\alpha) \). We will create a proposal density \( g \) by composition by splitting the domain at \( x = 1 \). Notice that the value of \( e^{-x} \) at \( x = 1 \) is \( 1/e \), but the value of \( x^{\alpha-1} \) at \( x = 1 \) is 1, which is \( e \) times as big. We need this observation to stitch together the components of \( g \).
If $W_1$ is distributed as $G(\alpha_1, 1)$ and $W_2$ as $G(\alpha_2, 1)$ and they are independent, then $W_1 + W_2$ is distributed as $G(\alpha_1 + \alpha_2, 1)$.

Decompose the event $W_1 + W_2 = w$ according to $W_1$; in infinitesimals we have

$$
\Pr(W_1 + W_2 = w) = \sum_{0 < x < w} \Pr(W_1 + W_2 = w \mid W_1 = x)\Pr(W_1 = x)
$$

$$
= \sum_{0 < x < w} \Pr(W_2 = w - x)\Pr(W_1 = x)
$$

$$
= \int_0^w \frac{(w - x)\alpha_2 - 1}{(\alpha_2 - 1)!} e^{-(w-x)} \frac{x^{\alpha_1 - 1}}{(\alpha_1 - 1)!} e^{-x} dx
$$

$$
= \frac{e^{-w}}{(\alpha_2 - 1)! (\alpha_1 - 1)!} \int_0^w (w - x)^{\alpha_2 - 1} x^{\alpha_1 - 1} dx.
$$

Suppose $\alpha_1$ is an integer; then use integration by parts, letting $u = x^{\alpha_1 - 1}$ and $dv = (w - x)^{\alpha_2 - 1} dx$. The integral becomes

$$
\int_0^w \frac{(w - x)\alpha_2 - 1 + 1}{(\alpha_2 - 1 + 1)} (\alpha_1 - 1)x^{\alpha_1 - 1 - 1} dx.
$$

Continue performing integration by parts, a total of $n$ times, until $\alpha_1 - 1 - n = 0$, i.e., $n = \alpha_1 - 1$ times. Then we will have

$$
\Pr(W_1 + W_2 = w) = \frac{e^{-w}}{(\alpha_2 - 1 + n)!} \int_0^w (w - x)^{\alpha_2 - 1 + n} dx
$$

$$
= \frac{w^{\alpha_2 + n} e^{-w}}{(\alpha_2 + n)!} = \frac{w^{\alpha_1 + \alpha_2 - 1} e^{-w}}{(\alpha_1 + \alpha_2 - 1)!}.
$$

**Fig. 2.14.** Derivation of mean and variance of the Poisson distribution.

Let $f_1$ be the distribution on $[0, 1]$ defined by

$$
f_1(x) = \alpha x^{\alpha - 1}, \quad \text{cdf}_1(x) = x^\alpha.
$$

And let $f_2$ be the distribution on $[1, \infty)$ defined by

$$
f_2(x) = e^{1-x}, \quad \text{cdf}_2(x) = 1 - e^{1-x}.
$$

Let $g$ be the composite density

$$
g(x) = p f_1(x) + (1 - p) f_2(x).
$$

It will be convenient to define $\chi_1(x)$ as the characteristic function of $[0, 1)$,
\[ \chi_1(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1, \\ 0 & \text{otherwise}, \end{cases} \]

and \[ \chi_2(x) \] as the characteristic function of \([1, \infty)\),

\[ \chi_2(x) = \begin{cases} 1 & \text{if } 1 \leq x, \\ 0 & \text{otherwise}. \end{cases} \]

Then \( g \) may be written

\[
g(x) = p\alpha x^{\alpha-1} \chi_1(x) + (1 - p)e^{1-x} \chi_2(x),
\]

where \( 0 < p < 1 \) is our choice.

For reasons that will be clear in a moment, choose \( p \) such that \( \lim_{x\uparrow 1} g(x) = e \lim_{x\downarrow 1} g(x) \). This gives

\[
p\alpha = e(1 - p), \quad \text{so } p = \frac{e}{e+\alpha} \text{ and } 1 - p = \frac{\alpha}{e+\alpha}.
\]

With this, \( g \) becomes

\[
g(x) = \frac{e\alpha}{e + \alpha} (x^{\alpha-1} \chi_1(x) + e^{-x} \chi_2(x)).
\]

Now we must find \( M \) such that \( Mg \geq f \).

For \( 0 \leq x \leq 1 \) the maximum value of \( e^{-x} \) is 1; hence choose \( M \) such that

\[
M \frac{e\alpha}{e + \alpha} x^{\alpha-1} = x^{\alpha-1} \geq x^{\alpha-1} e^{-x}.
\]

This gives

\[
M = \frac{e + \alpha}{e\alpha}.
\]

This value of \( M \) works for \( x \geq 1 \) as well; thus

\[
Mg = M \frac{e\alpha}{e + \alpha} e^{-x} = e^{-x} \geq x^{\alpha-1} e^{-x},
\]

where we note that the maximum value of \( x^{\alpha-1} \) is 1 for \( x \geq 1 \).

The acceptance discipline is

\[
\frac{f(Y)}{Mg(Y)} = \begin{cases} e^{-Y} & \text{on } [0, 1), \\ Y^{\alpha-1} & \text{on } [1, \infty). \end{cases}
\]

Altogether, the algorithm is:

1. let \( U_1, U_2 \sim U(0, 1) \),
2. if \( U_1 < \frac{e}{e+\alpha} \), case A, put \( Y = U_2^{1/\alpha} \); otherwise, case B, put \( Y = 1 - \ln U_2 \).
3. as the side experiment, let $R \sim U(0,1)$; in case A, return $X = Y$ if $R < e^{-Y}$; in case B, return $X = Y$ if $R < Y^{\alpha - 1}$; otherwise go to step 1.

The efficiency of the algorithm is

$$1/M = \frac{e\alpha}{e + \alpha}.$$

Now let $\alpha > 0$ be arbitrary and let $m = \text{int} [\alpha]$ and $\delta = \text{frac}(\alpha)$. By the reproductive property, if $Y \sim G(m, 1)$ and $Z \sim G(\delta, 1)$, then $X = Y + Z \sim G(\alpha, 1)$. So it remains to show how to generate samples from the Erlang distribution $G(m, 1)$. But this is easily accomplished by simulation. In Section 2.5.1 we saw that $W = -\ln U$, where $U \sim U(0,1)$ is a sample from $E(1)$; therefore $-\sum_{1}^{m} \ln(U_i), U_i \sim U(0,1)$, is a sample from $G(m, 1)$. This may be simplified slightly:

1. for $i = 1, 2, \ldots, m$ let $U_i \sim U(0,1)$;
2. return $Y = -\ln(\prod_{i=1}^{m} U_i)$.

The returned value $Y$ is distributed like $G(m, 1)$.

**Notes**

The problem of devising algorithms for sampling from various distributions has attracted considerable interest. The book [Rub81] has a wealth of algorithms for many distributions. Probably the biggest effort in this regard is the two-volume treatise published by the Los Alamos National Laboratory based on work done at the lab shortly after the Monte Carlo method was invented.

There is also a large interest in obtaining random samples of various types of combinatorial objects such as graphs, permutations, partitions, and tilings. The Markov chain Monte Carlo method (see Chapter 3) has both been greatly influenced by and been very useful in this search.

Multivariate distributions are very common, and thus it is a common problem to try to extract structure from data derived from a multivariate distribution. This is the central problem in the area of data mining, and many different ways have been devised to graphically investigate high-dimensional data sets. Clearly there is no one correct way of doing this.

**Problems: Chapter 2**

1. (4) (St. Petersburg problem) (a) Player A pays house B $5 and flips a coin. If it falls tails, the game ends. If it falls heads, B gives A $1 and A flips the coin again. This continues until the coin lands tails, with B’s
payoff to A doubling for each successive head. Simulate this game 100 times, then 1000 times, then 10,000 times and histogram A’s winnings. (b) Keep track of A’s fortune during the play of the series of games. What observations can you make about this situation?

2. (5) (Paris Salon problem) (a) Let $X$ be the random variable for the first time a 6 is obtained on the roll of a fair die. This should take 6 rolls on average, should it not? Simulate this experiment (many times) to obtain a histogram and to estimate the expectation of $X$. (b) Given the results of (a) would you bet $1 to make $1 that a 6 will not be obtained in 4 rolls of a fair die? Simulate this experiment and report.

3. (3) Let $g$ be the piecewise constant density

$$g(x) = \begin{cases} c, & 0 \leq x < 0.1, \\ 2c, & 0.1 \leq x < 0.7, \\ c/3, & 0.7 \leq x < 1.0, \\ 0 & \text{elsewhere}. \end{cases}$$

and 0 elsewhere. Find $c$ such that this is a probability density function. Show how to sample from $g$ by inverting the cdf.

4. (4) (Composite density) Let $f$ be the piecewise linear function given by

$$f(x) = \begin{cases} x/4, & 0 \leq x < 1, \\ 1/4, & 1 \leq x < 3, \\ (x - 3)/60 - (x - 6)/12, & 3 \leq x < 6, \\ 1/20, & 6 \leq x < 15, \\ 0 & \text{elsewhere}. \end{cases}$$

and 0 elsewhere. Find the constant $c$ such that $g(x) = cf(x)$ is a probability density. What are the respective probabilities that a sample falls within the regions A $[0, 1)$, B $[1, 3)$, C $[3, 6)$, and D $[6, 15]$? Devise a method for sampling from $g$ and implement it.

5. (4) Let $g$ be the function on $[-2, 2]$ given by

$$g(x) = \frac{8}{7} + \frac{118}{63} x^2 - \frac{74}{63} x^4 + \frac{10}{63} x^6.$$ 

Find the constant $c$ such that $f(x) = cg(x)$ is the pdf for a distribution. Show how to sample from this density.

6. (5) (Points in a box) $N$ points are placed uniformly at random in the unit square. What is the expected minimum distance between them? Simulate the problem for different values of $N$ and graph the distance versus $N$.

7. (4) (Sampling from a multivariate normal) If $X_1, X_2 \sim N(0, 1)$, argue that the vector $X = (X_1, X_2)$ has a distribution whose pdf has circular symmetry. In fact, the pdf is the function (where $C$ is a normalizing constant)
Problems: Chapter 2 95

\[ f(x_1, x_2) = Ce^{-(x_1^2 + x_2^2)/2}. \]

This is an example of a \textit{bivariate normal distribution}. Now, given a \(2 \times 2\) matrix \(A\) and a vector \(\mu = (\mu_1, \mu_2)\), let

\[ (Y_1, Y_2) = Y = AX + \mu \]

and investigate the distribution of \(Y\).

8. (4) (Random point from a sphere) The objective of this problem is to devise a way to find a random sample from the sphere

\[ S = \{(x_1, x_2, \ldots, x_n) : x_1^2 + x_2^2 + \cdots + x_n^2 = 1\} \]

in \(\mathbb{R}^n\). The hint to do this is to find a multivariate distribution in \(\mathbb{R}^n\) whose pdf \(f\) has spherical symmetry (that is, \(f(x) = f(y)\) if \(x\) and \(y\) are the same distance from the origin; see Problem 7). Then sample a point \((X_1, X_2, \ldots, X_n)\) from \(f\), and the required sample should be

\[ \frac{(X_1, X_2, X_3, \ldots, X_n)}{(X_1^2 + X_2^2 + \cdots + X_n^2)^{1/2}}. \]

Try out your solution in \(\mathbb{R}^2\) (obtaining points on the unit circle) and histogram them along the line segment \([0, 2\pi]\).

9. (4) (Distribution of the minimum) Fix \(N = 10\), for the moment. Choose independently \(X_i \sim U(0, 1)\) for \(i = 1, 2, \ldots, 10\); the goal is to find the distribution of min \(X_i\). That is, we want to know \(\Pr(\text{min} X_i \leq t)\) for each \(t \in [0, 1]\). Do this by simulation, that is, generate multiple trials of this experiment and histogram the random variable \(\text{min} X_i\). How do the results depend on \(N\)? To see this, change \(N\) and repeat.

10. (4) (Distribution of the minimum) Assuming that the \(X_i\) are independent, argue that

\[ \Pr(\text{min} X_i > t) = \prod_i \Pr(X_i > t) \]

to compute the pdf and cdf of the distribution of min \(X_i\), where \(X_i \sim U(0, 1)\).

11. (4) Show that if \(X\) and \(Y\) are independently distributed with cdfs \(F_X\) and \(F_Y\) respectively, then \(Z = \max\{X, Y\}\) has cdf

\[ F_Z(z) = F_X(z)F_Y(z). \]

Use this to find an algorithm to sample from

\[ Z = \max\{U_1, U_2, \ldots, U_n\}, \]

where each \(U_j \sim U(0, 1)\).
12. (3) Obtain samples from a Poisson random variable with parameter \( \lambda = 0.1 \) per minute for \( t = 30 \) minutes by sampling from (a) \( b(20, \frac{\lambda}{20}) \), (b) \( b(60, \frac{\lambda}{60}) \), and (c) \( b(100, \frac{\lambda}{100}) \). Histogram the results of all three and compare. Keep track of run times.

13. (3) Obtain samples from a Poisson random variable with parameter \( \lambda = 0.1 \) by “Method 2” from Section 2.5.1 and histogram the results. Compare with the results from the previous problem.

14. (4) (Sampling bias for bus waiting times) Suppose the interarrival time for a city bus has an exponential distribution with parameter \( 1/\lambda \). A passenger arrives at a uniformly random time and records the time until the next bus arrives. What is the expected waiting time? Use a simulation to get an answer. Is the answer surprising? Now suppose instead that the interarrival time is \( U(0, 2\lambda) \). How does this change the situation? (Notice that the expected interarrival time is \( \lambda \) in both cases.)

15. (2) Devise an algorithm to obtain samples from the distribution on \([0, 1]\) with pdf \( f(x) = nx^{n-1} \) for any fixed natural number \( n \).

16. (5) (Difference between equally likely and conditional equally likely; see [Ross96] page 67) This problem examines a situation in which exponential waiting times yield an “equally likely” distribution. Consider an exponential distribution with parameter \( \lambda \) and a situation in which exactly \( n = 50 \) events occurred in the interval \([0, 1]\). Compare this situation to one obtained by choosing 50 samples \( U_i \) from \( U(0, 1) \) as the events. For one side of the comparison, let \( U_i \) be 50 samples from \( U(0, 1) \). For example, suppose the random numbers are

\[
0.48, 0.36, 0.66, 0.12, 0.82, 0.57, 0.28, 0.11, 0.18, 0.71.
\]

Now sort them to get

\[
0.11, 0.12, 0.18, 0.28, 0.36, 0.48, 0.57, 0.66, 0.71, 0.82,
\]

and histogram the interpoint distances

\[
(0.12 - 0.11 =) 0.01, \quad (0.18 - 0.12 =) 0.06, \quad (0.28 - 0.18 =) 0.10, \\
(0.36 - 0.28 =) 0.08, \quad (0.48 - 0.36 =) 0.08, \quad (0.57 - 0.66 =) 0.11, \\
(0.66 - 0.71 =) 0.04, \quad (0.82 - 0.71 =) 0.11.
\]

To get an estimate of the distribution, this needs to be done many times. For the other side of the comparison it is necessary to generate \( E(\lambda) \) waiting times that yield exactly 50 events in the interval \([0, 1]\). The parameter \( \lambda \) should be set so that the expected number of events is \( n = 50 \). The simplest (but certainly not the most efficient) method to do this is by
rejection, that is, generate samples from $E(\lambda)$ and check whether exactly 50 fit in the interval $[0, 1]$. How inefficient is this? That is, how many rejections does one get per acceptance? Can you come up with another method?

17. (3) Show that the following “zombie gap” algorithm approximately simulates exponential arrivals. Find the exponential parameter. (From Scientific American Computer Recreations, April 1985.) Let the “gap” be from 0 to $w$, $w < 1$.

```
loop:
count = 0
loop:
    increment count
    choose $U \sim U(0, 1)$
    if($U < w$) break out of loop
endloop
report count as next arrival time
endloop
```

Here is a sample run: 2, 4, 1, 1, 3, 1, 2, 5, ..., meaning that the first event occurred at time 2, the next at time 6 (interarrival time of 4), etc. Show that the interarrival times are approximately exponentially distributed, especially for small $w$. What is the exponential parameter?

18. (3) Suppose a species of bacteria reproduce with exponentially distributed waiting times having parameter $\lambda = 2$. Further suppose there is a “survival reward” associated with a birth at time $t$ given by $r(t) = te^{-t}$. By simulation, estimate the expectation of $r$. What is the exact value? (Hint, \[
\int_0^\infty te^{-t}\lambda e^{-\lambda t} dt = (\lambda/(\lambda + 1)^2) = 2/9.\]

19. (6) (Pure birth process.) Suppose a newly born bacterium waits an exponentially distributed period of time $T$ with rate parameter $\lambda$, in short $T \sim E(\lambda)$, and then divides, replacing itself with two newly born bacteria. Simulate the growth of such a colony starting from a single bacterium for 60 minutes assuming $\lambda = 0.05$ per minute. What is the (empirical) doubling time for such a colony? What is the average number of generations in this period of time (60 minutes)? (Note: one can derive the differential equations for a birth process exactly as in the Poisson derivation, the only difference being that at the time there are $k$ cells, then the event rate parameter is $k\lambda$ instead of just $\lambda$; hence the system of differential equations becomes \[
dP_k(t)/dt = -k\lambda P_k(t) + \lambda(k - 1)P_{k-1}(t),\]
and their solution is $P_k(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{k-1}$, $k = 1, 2, \ldots$, where $P_k(t) = \text{Pr}(X(t) = k)$. The expectation is $\mathbb{E}(X(t)) = e^{\lambda t}$, i.e., exponential growth. Hence the doubling time for a colony of bacteria is the solution to $\ln(2) = \lambda t$, while for a single bacterium, it is the solution to $1 = \lambda t$ via the Poisson expectation
\[ \mathbb{E}(X(t)) = \lambda t. \]

20. (5) If \( T_\lambda \sim E(\lambda) \) and, independently, \( T_\mu \sim E(\mu) \), what is the probability that \( T_\lambda < T_\mu \)? (Hint, the joint pdf is \( \lambda \mu e^{-\lambda r} e^{-\mu s} \).)

21. (7) (Birth and death process.) When a cell is “born,” draw a sample \( T_b \sim E(\lambda) \) and a sample \( T_d \sim E(\mu) \). If \( T_b \leq T_d \), then the simulated cell divides at time \( T_b \) into two new cells as above (and \( T_d \) is discarded). If \( T_d < T_b \), then the cell dies (and \( T_b \) is discarded). Simulate this process for 20 minutes starting from a single cell with \( \mu = 1 \) per minute and (a) \( \lambda = 1 \), (b) \( \lambda = 1.05 \), and (c) \( \lambda = 1.10 \) all in units of per minute. Compute the empirical mean and variance of the number of cells at the end of the study.

22. (7) (Single-server queue.) People join a queue according to the exponential distribution \( E(\lambda) \). The service time for the person at the front of the queue is exponential \( E(\mu) \). Simulate the process for 30 minutes for several runs. Experiment with different values of \( \lambda \) and \( \mu \); note that if \( \mu \) is not close to \( \lambda \), then the queue grows exponentially. Graph the length as a function of time for one of the runs. Find the mean and variance of the waiting time in the queue.

23. (7) (Lanchester battlefield model.) Let \( B \) denote the number of blue soldiers and \( R \) the number of red soldiers. Blue soldiers have killing effectiveness parameter \( \mu_B \) and red soldiers have effectiveness \( \mu_R \). The Blue army starts with \( B_0 \) soldiers and the Red army with \( R_0 \). The waiting time for the next death of a Blue soldier is exponential \( E(R\mu_R) \), where \( R \) is the number of Red soldiers at that time and the waiting time for the death of a Red soldier is similarly exponential \( E(B\mu_B) \). To start, sample \( T_B \sim E(R_0\mu_R) \) and \( T_R \sim E(B_0\mu_B) \). Move time forward to each event, handle it, and schedule the next. Thus, if \( T_B < T_R \), move time forward to \( T_B \), decrement \( B \) by one, and resample \( T_B \sim E(R\mu_R) \). And continue until one army or the other is wiped out. Fix \( \mu_R = 1 \) and \( R_0 = 100 \) and experiment with values of \( \mu_B \) and \( B_0 \). In each case find the (mean) time of the battle and each side’s expectation of winning. Plot the battle strength vs time for both armies for a typical run.

24. (5) (Retirement benefit projection) At age 50 Fannie Mae has $150,000 invested and will be investing another $10,000 per year until age 70. Each year the investment grows according to an interest rate that is normally distributed with mean 8% and standard deviation 9%. At age 70, Fannie Mae then retires and withdraws $65,000 per year until death. Below is given a conditional death probability table. Thus if Fannie Mae lives until age 70, then the probability of dying before age 71 is 0.04979. Simulate this process 1000 times and histogram the amount of money Fannie Mae has at death.
<table>
<thead>
<tr>
<th>Age</th>
<th>Probability of dying during the year</th>
<th>Age</th>
<th>Probability of dying during the year</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.00832</td>
<td>64</td>
<td>0.02904</td>
</tr>
<tr>
<td>51</td>
<td>0.00911</td>
<td>65</td>
<td>0.03175</td>
</tr>
<tr>
<td>52</td>
<td>0.00996</td>
<td>66</td>
<td>0.03474</td>
</tr>
<tr>
<td>53</td>
<td>0.01089</td>
<td>67</td>
<td>0.03804</td>
</tr>
<tr>
<td>54</td>
<td>0.01190</td>
<td>68</td>
<td>0.04168</td>
</tr>
<tr>
<td>55</td>
<td>0.01300</td>
<td>69</td>
<td>0.04561</td>
</tr>
<tr>
<td>56</td>
<td>0.01421</td>
<td>70</td>
<td>0.04979</td>
</tr>
<tr>
<td>57</td>
<td>0.01554</td>
<td>71</td>
<td>0.05415</td>
</tr>
<tr>
<td>58</td>
<td>0.01700</td>
<td>72</td>
<td>0.05865</td>
</tr>
<tr>
<td>59</td>
<td>0.01859</td>
<td>73</td>
<td>0.06326</td>
</tr>
<tr>
<td>60</td>
<td>0.02034</td>
<td>74</td>
<td>0.06812</td>
</tr>
<tr>
<td>61</td>
<td>0.02224</td>
<td>75</td>
<td>0.07337</td>
</tr>
<tr>
<td>62</td>
<td>0.02431</td>
<td>76</td>
<td>0.07918</td>
</tr>
<tr>
<td>63</td>
<td>0.02657</td>
<td>77</td>
<td>0.08570</td>
</tr>
</tbody>
</table>

* Source: Society of Actuaries, Life Contingencies.

25. (4) (Median of exponential vs mean) Suppose the lifetime of lightbulbs is exponential with mean 1000 hours. Given a randomly selected lightbulb, what is the probability that it will still be burning after 1000 hours? Perform this experiment 100 times and histogram the event “less than 693 hours” versus “more than 693 hours.” (The median \( m \) is the point where \( \Pr(T \leq m) = \Pr(T > m) \), so \( m = \text{cdf}^{-1}(1/2) \). For the exponential, \( m = \ln(2)/\lambda = \mu \ln 2 \).)

26. (4) Implement all three of the normal sampling routines (via the central limit theorem, Box–Muller algorithm, Marsaglia–Bray algorithm). Test them for (a) speed, (b) accuracy (make a pdf and cdf for all three). Compare the results and give some conclusions.

27. (3) One way to check an RNG is to make a correlation plot. This is a plot of sequential pairs \((X_n, X_{n+1})\) generated by the RNG. Ideally, there should be no structure to this plot. Make a correlation plot for a sequence of normal pseudorandom variables using either the Box–Muller or Marsaglia–Bray algorithm.

28. (4) Test the central limit theorem by sampling (a) 10 times, (b) 100 times from an exponential distribution; form the normalized random variable; and plot a pdf and cdf. Compare with table values for the cdf, see Table 2.2. Repeat with several different values of the parameter \( \lambda \). How does \( \lambda \) influence the outcome?

29. (6) Repeat the previous problem except use several exponential, gamma, uniform, Poisson, and beta distributions. How does the choice of distribution affect the outcome? How does the choice of parameter for the
distribution effect the outcome? Does it make much difference whether the distribution is symmetric or not?

30. (3) Show that if $X \sim N(0, 1)$, then $Y = \sigma X + \mu$ is distributed as normal with mean $\mu$ and variance $\sigma^2$.

31. (5) (Lognormal distribution.) Let $X$ be normally distributed with mean $\alpha$ and variance $\beta$ and let $X = \ln Y$ (that is, $Y = e^X$). (a) Find the probability density function of $Y$. (b) Devise a way to sample from this distribution, i.e., to produce random variables $Y$ with the correct distribution.

32. (4) Let $g$ be the density $g(x) = (x - 2)/12 - (x - 6)/24$, $2 \leq x < 6$, and 0 elsewhere. Sample from $g$ by (a) inverting the cdf and (b) rejection. Histogram and compare the speed of drawing $N$ samples for a convenient $N$ (100,000 or 1,000,000 perhaps).

33. (3) Adapt the rejection method to sample from a beta distribution whose pdf is $f(x) = cx^{\alpha-1}(1-x)^{\beta-1}$, $\alpha > 0$, $\beta > 0$, and $0 \leq x \leq 1$. The constant $c$ is chosen so that $\int_0^1 f(x)dx = 1$. Use $\alpha = 5$ and $\beta = 2$, for example. Notice that the value of $c$ is not needed (except to compute the optimal value for $M$). Make a histogram.

34. (5) Devise an envelope function $g$ to use in the rejection method to sample from the distribution whose pdf is $f(x) = cx^2e^{-x}$, $0 \leq x < \infty$, where $c$ is chosen so that $\int_0^\infty f(x)dx = 1$. Test your samples for accuracy with the pdf by histogramming, check for correlation (by graphing $(x_n, x_{n+1})$), and test for efficiency (i.e., rate of generation) and the effect of the starting point (by doing a pdf starting from different points).

35. (6) Do a pure birth model, see Exercise 19 above, using a gamma distribution in place of the exponential distribution for waiting times.

36. (7) Do a birth and death model, see Exercise 21 above, using a gamma distribution in place of the exponential distribution for waiting times.

37. (4) Can you think of an alternate way of obtaining a probability $p = 1/3$ Bernoulli trial random variable (or any other for $0 < p < 1$) using coin flips? (Hint: (thanks to Michael Lacy) express $p$ in binary.)
Explorations in Monte Carlo Methods
Shonkwiler, R.W.; Mendivil, F.
2009, XII, 243 p., Hardcover