the one with the more variable service times (larger σ^2) will tend to have longer lines on the average. Intuitively, if service times are highly variable, there is a high probability that a large service time will occur (say, much larger than the mean service time), and, when large service times do occur, there is a higher-than-usual tendency for lines to form and delays of customers to increase. (The reader should not confuse "steady state" with low variability or short lines; a system in steady-state or statistical equilibrium can be highly variable and can have long waiting lines.)

Example 6.9

There are two workers competing for a job. Able claims an average service time that is faster than Baker's, but Baker claims to be more consistent, even if not as fast. The arrivals occur according to a Poisson process at the rate $\lambda = 2$ per hour (1/30 per minute). Able's service statistics are an average service time of 24 minutes with a standard deviation of 20 minutes. Baker's service statistics are an average service time of 25 minutes, but a standard deviation of only 2 minutes. If the average length of the queue is the criterion for hiring, which worker should be hired? For Able, $\lambda = 1/30$ per minute, $1/\mu = 24$ minutes, $\sigma^2 = 20^2 = 400$ minutes^2, $\rho = \lambda/\mu = 24/30 = 4/5$, and the average queue length is computed as

$$L_Q = \frac{(1/30)^2[24^2 + 400]}{2(1 - 4/5)} = 2.711 \text{ customers}$$

For Baker, $\lambda = 1/30$ per minute, $1/\mu = 25$ minutes, $\sigma^2 = 2^2 = 4$ minutes^2, $\rho = 25/30 = 5/6$, and the average queue length is

$$L_Q = \frac{(1/30)^2[25^2 + 4]}{2(1 - 5/6)} = 2.097 \text{ customers}$$

Although working faster on the average, Able's greater service variability results in an average queue length about 30% greater than Baker's. On the basis of average queue length, L_Q, Baker wins. On the other hand, the proportion of arrivals who would find Able idle and thus experience no delay is $P_0 = 1 - \rho = 1/5 = 20\%$, but the proportion who would find Baker idle and thus experience no delay is $P_0 = 1 - \rho = 1/6 = 16.7\%$.

One case of the $M/G/1$ queue that is of special note occurs when service times are exponential, which we describe next.

The M/M/1 queue. Suppose that service times in an $M/G/1$ queue are exponentially distributed, with mean $1/\mu$; then the variance as given by Equation (5.27) is $\sigma^2 = 1/\mu^2$. The mean and standard deviation of the exponential distribution are equal, so the $M/M/1$ queue will often be a useful approximate model when service times have standard deviations approximately equal to their means. The steady-state parameters, given in Table 6.4, may be computed by substituting $\sigma^2 = 1/\mu^2$ into the formulas in Table 6.3. Alternatively, L may be computed by Equation (6.16) from the steady-state probabilities P_n given in Table 6.4, and then w, w_Q, and L_Q may be computed from Equations (6.17). The student can show that the two expressions for each parameter are equivalent by substituting $\rho = \lambda/\mu$ into the right-hand side of each equation in Table 6.4.

Example 6.10

Suppose that the interarrival times and service times at a single-chair unisex hair-styling shop have been shown to be exponentially distributed. The values of λ and μ are 2 per hour and 3 per hour, respectively—that is, the time between arrivals averages 1/2 hour, exponentially distributed, and the service time averages 20 minutes, also exponentially distributed. The server utilization and the probabilities for zero, one, two, three, and four or more customers in the shop are computed as follows:

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3}$$

$$P_0 = 1 - \frac{\lambda}{\mu} = \frac{1}{3}$$

$$P_1 = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{2}{9}$$

$$P_2 = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^2 = \frac{4}{27}$$

$$P_3 = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^3 = \frac{8}{81}$$

$$P_{\geq 4} = 1 - \sum_{n=0}^{3} P_n = 1 - \frac{1}{3} - \frac{2}{9} - \frac{4}{27} - \frac{8}{81} = \frac{16}{81}$$

From the calculations, the probability that the hair stylist is busy is $1 - P_0 = \rho = 0.67$; thus, the probability that the hair stylist is idle is 0.33. The time-average number of customers in the system is given by Table 6.4 as

$$L = \frac{\lambda}{\mu - \lambda} = \frac{2}{3-2} = 2 \text{ customers}$$

The average time an arrival spends in the system can be obtained from Table 6.4 or Equation (6.17) as

$$w = \frac{L}{\lambda} = \frac{2}{2} = 1 \text{ hour}$$

The average time the customer spends in the queue can be obtained from Equation (6.17) as

$$w_Q = w - \frac{1}{\mu} = 1 - \frac{1}{3} = \frac{2}{3} \text{ hour}$$

Table 6.4 Steady-State Parameters of the $M/M/1$ Queue

L	$\dfrac{\lambda}{\mu-\lambda} = \dfrac{\rho}{1-\rho}$
w	$\dfrac{1}{\mu-\lambda} = \dfrac{1}{\mu(1-\rho)}$
w_Q	$\dfrac{\lambda}{\mu(\mu-\lambda)} = \dfrac{\rho}{\mu(1-\rho)}$
L_Q	$\dfrac{\lambda^2}{\mu(\mu-\lambda)} = \dfrac{\rho^2}{1-\rho}$
P_n	$\left(1-\dfrac{\lambda}{\mu}\right)\left(\dfrac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n$

From Table 6.4, the time-average number in the queue is given by

$$L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{4}{3(1)} = \frac{4}{3} \text{ customers}$$

Finally, notice that multiplying $w = w_Q + 1/\mu$ through by λ and using Little's equation (6.9) yields

$$L = L_Q + \frac{\lambda}{\mu} = \frac{4}{3} + \frac{2}{3} = 2 \text{ customers}$$

Example 6.11

For the $M/M/1$ queue with service rate $\mu = 10$ customers per hour, consider how L and w increase as the arrival rate, λ, increases from 5 to 8.64 by increments of 20%, and then to $\lambda = 10$.

λ	5.0	6.0	7.2	8.64	10.0
ρ	0.500	0.600	0.720	0.864	1.0
L	1.00	1.50	2.57	6.35	∞
w	0.20	0.25	0.36	0.73	∞

For any $M/G/1$ queue, if $\lambda/\mu \geq 1$, waiting lines tend to continually grow in length; the long-run measures of performance, L, w, w_Q, and L_Q are all infinite ($L = w = w_Q = L_Q = \infty$); and a steady-state probability distribution does not exist. As is shown here for $\lambda < \mu$, if ρ is close to 1, waiting lines and delays will tend to be long. Notice that the increase in average system time, w, and average number in system, L, is highly nonlinear as a function of ρ. For example, as λ increases by 20%, L increases first by 50% (from 1.00 to 1.50), then by 71% (to 2.57), and then by 147% (to 6.35).

Example 6.12

If arrivals are occurring at rate $\lambda = 10$ per hour, and management has a choice of two servers, one who works at rate $\mu_1 = 11$ customers per hour and the second at rate $\mu_2 = 12$ customers per hour, the respective utilizations are $\rho_1 = \lambda/\mu_1 = 10/11 = 0.909$ and $\rho_2 = \lambda/\mu_2 = 10/12 = 0.833$. If the $M/M/1$ queue is used as an approximate model, then, with the first server, the average number in the system would be, by Table 6.4,

$$L_1 = \frac{\rho_1}{1-\rho_1} = 10$$

and, with the second server, the average number in the system would be

$$L_2 = \frac{\rho_2}{1-\rho_2} = 5$$

Thus, a decrease in service rate from 12 to 11 customers per hour, a mere 8.3% decrease, would result in an increase in average number in system from 5 to 10, which is a 100% increase.

The effect of utilization and service variability

For any $M/G/1$ queue, if lines are too long, they can be reduced by decreasing the server utilization ρ or by decreasing the service time variability, σ^2. These remarks hold for almost all queues, not just the $M/G/1$ queue. The utilization factor ρ can be reduced by decreasing the arrival rate λ, by increasing the service rate μ, or by increasing the number of servers, because, in general, $\rho = \lambda/(c\mu)$, where c is the number of parallel servers. The effect of additional servers will be studied in the following subsections. Figure 6.12 illustrates the effect of service variability. The mean steady-state number in the queue, L_Q, is plotted versus utilization

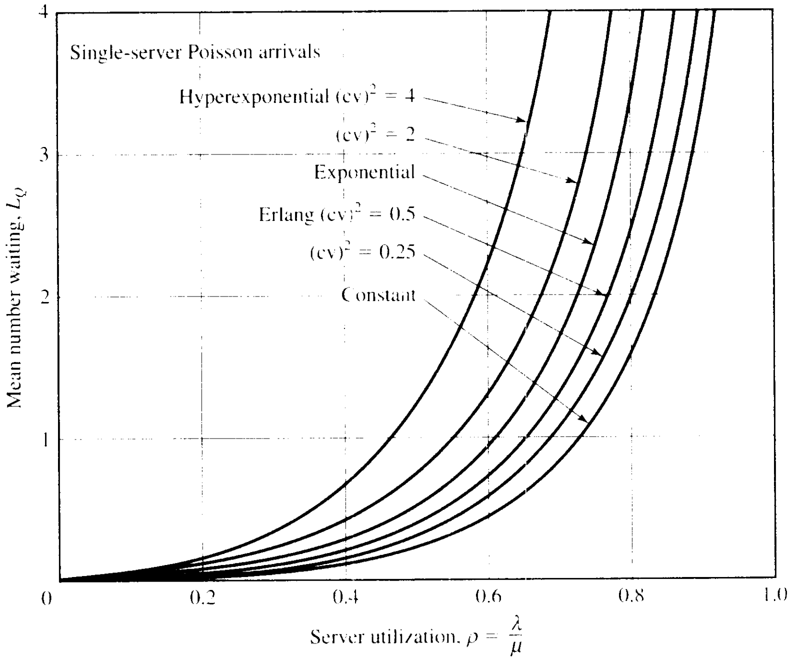

Figure 6.12 Mean number of customers waiting, L_Q, in $M/G/1$ queue having service distributions with given cv. (Adapted from Geoffrey Gordon, *System Simulation*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1978.)

ρ for a number of different coefficients of variation. The coefficient of variation (cv) of a positive random variable X is defined as

$$(\text{cv})^2 = \frac{V(X)}{[E(X)]^2}$$

It is a measure of the variability of a distribution. The larger its value, the more variable is the distribution relative to its expected value. For deterministic service times, $V(X) = 0$, so cv = 0. For Erlang service times of order k, $V(X) = 1/(k\mu^2)$ and $E(X) = 1/\mu$, so $\text{cv} = 1/\sqrt{k}$. For exponential service times at service rate μ, the mean service time is $E(X) = 1/\mu$ and the variance is $V(X) = 1/\mu^2$, so cv = 1. If service times have standard deviation greater than their mean (i.e., if cv > 1), then the hyperexponential distribution, which can achieve any desired coefficient of variation greater than 1, provides a good model. One occasion where it arises is given in Exercise 16.

The formula for L_Q for any $M/G/1$ queue can be rewritten in terms of the coefficient of variation by noticing that $(\text{cv})^2 = \sigma^2/(1/\mu)^2 = \sigma^2\mu^2$. Therefore,

$$\begin{aligned} L_Q &= \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} \\ &= \frac{\rho^2(1+(\text{cv})^2)}{2(1-\rho)} \\ &= \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{1+(\text{cv})^2}{2}\right) \end{aligned} \qquad (6.18)$$

The first term, $\rho^2/(1-\rho)$, is L_Q for an $M/M/1$ queue. The second term, $(1 + (\text{cv})^2)/2$, corrects the $M/M/1$ formula to account for a nonexponential service-time distribution. The formula for w_Q can be obtained from the corresponding $M/M/1$ formula by applying the same correction factor.

6.4.2 Multiserver Queue: M/M/c/∞/∞

Suppose that there are c channels operating in parallel. Each of these channels has an independent and identical exponential service-time distribution, with mean $1/\mu$. The arrival process is Poisson with rate λ. Arrivals will join a single queue and enter the first available service channel. The queueing system is shown in Figure 6.13. If the number in system is $n < c$, an arrival will enter an available channel. However, when $n \geq c$, a queue will build if arrivals occur.

The offered load is defined by λ/μ. If $\lambda \geq c\mu$, the arrival rate is greater than or equal to the maximum service rate of the system (the service rate when all servers are busy); thus, the system cannot handle the load put upon it, and therefore it has no statistical equilibrium. If $\lambda > c\mu$, the waiting line grows in length at the rate $(\lambda - c\mu)$ customers per time unit, on the average. Customers are entering the system at rate λ per time unit but are leaving the system at a maximum rate of $c\mu$ per time unit.

For the $M/M/c$ queue to have statistical equilibrium, the offered load must satisfy $\lambda/\mu < c$, in which case $\lambda/(c\mu) = \rho$, the server utilization. The steady-state parameters are listed in Table 6.5. Most of the measures of performance can be expressed fairly simply in terms of P_0, the probability that the system is empty, or $\sum_{n=c}^{\infty} P_n$, the probability that all servers are busy, denoted by $P(L(\infty) \geq c)$, where $L(\infty)$ is a random variable representing the number in system in statistical equilibrium (after a very long time). Thus, $P(L(\infty) = n) = P_n$, $n = 0, 1, 2, \ldots$. The value of P_0 is necessary for computing all the measures of performance, and the equation for P_0 is somewhat more complex than in the previous cases. However, P_0 depends only on c and ρ. A good approximation to P_0 can be obtained by using Figure 6.14, where P_0 is plotted versus ρ on semilog paper for various values c. Figure 6.15 is a plot of L versus ρ for different values of c.

The results in Table 6.5 simplify to those in Table 6.4 when $c = 1$, the case of a single server. Notice that the average number of busy servers, or the average number of customers being served, is given by the simple expression $L - L_Q = \lambda/\mu = c\rho$.

Example 6.13

Many early examples of queueing theory applied to practical problems concerning tool cribs. Attendants manage the tool cribs as mechanics, assumed to be from an infinite calling population, arrive for service. Assume Poisson arrivals at rate 2 mechanics per minute and exponentially distributed service times with mean 40 seconds.

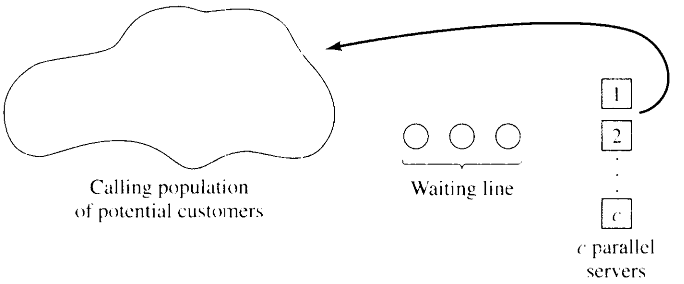

Figure 6.13 Multiserver queueing system.

Table 6.5 Steady-State parameters for the $M/M/c$ Queue

ρ	$\dfrac{\lambda}{c\mu}$
P_0	$\left\{\left[\sum_{n=0}^{c-1}\dfrac{(\lambda/\mu)^n}{n!}\right]+\left[\left(\dfrac{\lambda}{\mu}\right)^c\left(\dfrac{1}{c!}\right)\left(\dfrac{c\mu}{c\mu-\lambda}\right)\right]\right\}^{-1}$
	$=\left\{\left[\sum_{n=0}^{c-1}\dfrac{(c\rho)^n}{n!}\right]+\left[(c\rho)^c\left(\dfrac{1}{c!}\right)\left(\dfrac{1}{1-\rho}\right)\right]\right\}^{-1}$
$P(L(\infty)\geq c)$	$\dfrac{(\lambda/\mu)^c P_0}{c!(1-\lambda/c\mu)}=\dfrac{(c\rho)^c P_0}{c!(1-\rho)}$
L	$c\rho+\dfrac{(c\rho)^{c+1}P_0}{c(c!)(1-\rho)^2}=c\rho+\dfrac{\rho P(L(\infty)\geq c)}{(1-\rho)}$
w	$\dfrac{L}{\lambda}$
w_Q	$w-\dfrac{1}{\mu}$
L_Q	$\lambda w_Q=\dfrac{(c\rho)^{c+1}P_0}{c(c!)(1-\rho)^2}=\dfrac{\rho P(L(\infty)\geq c)}{(1-\rho)}$
$L-L_Q$	$\dfrac{\lambda}{\mu}=c\rho$

Now, $\lambda = 2$ per minute, and $\mu = 60/40 = 3/2$ per minute. The offered load is greater than 1:

$$\frac{\lambda}{\mu}=\frac{2}{3/2}=\frac{4}{3}>1$$

so more than one server is needed if the system is to have a statistical equilibrium. The requirement for steady state is that $c > \lambda/\mu = 4/3$. Thus at least $c = 2$ attendants are needed. The quantity 4/3 is the expected number of busy servers, and for $c \geq 2$, $\rho = 4/(3c)$ is the long-run proportion of time each server is busy. (What would happen if there were only $c = 1$ server?)

Let there be $c = 2$ attendants. First, P_0 is calculated as

$$P_0=\left\{\sum_{n=0}^{1}\frac{(4/3)^n}{n!}+\left(\frac{4}{3}\right)^2\left(\frac{1}{2!}\right)\left[\frac{2(3/2)}{2(3/2)-2}\right]\right\}^{-1}$$

$$=\left\{1+\frac{4}{3}+\left(\frac{16}{9}\right)\left(\frac{1}{2}\right)(3)\right\}^{-1}=\left(\frac{15}{3}\right)^{-1}=\frac{1}{5}=0.2$$

Next, the probability that all servers are busy is computed as

$$P(L(\infty)\geq 2)=\frac{(4/3)^2}{2!(1-2/3)}\left(\frac{1}{5}\right)=\left(\frac{8}{3}\right)\left(\frac{1}{5}\right)=\frac{8}{15}=0.533$$

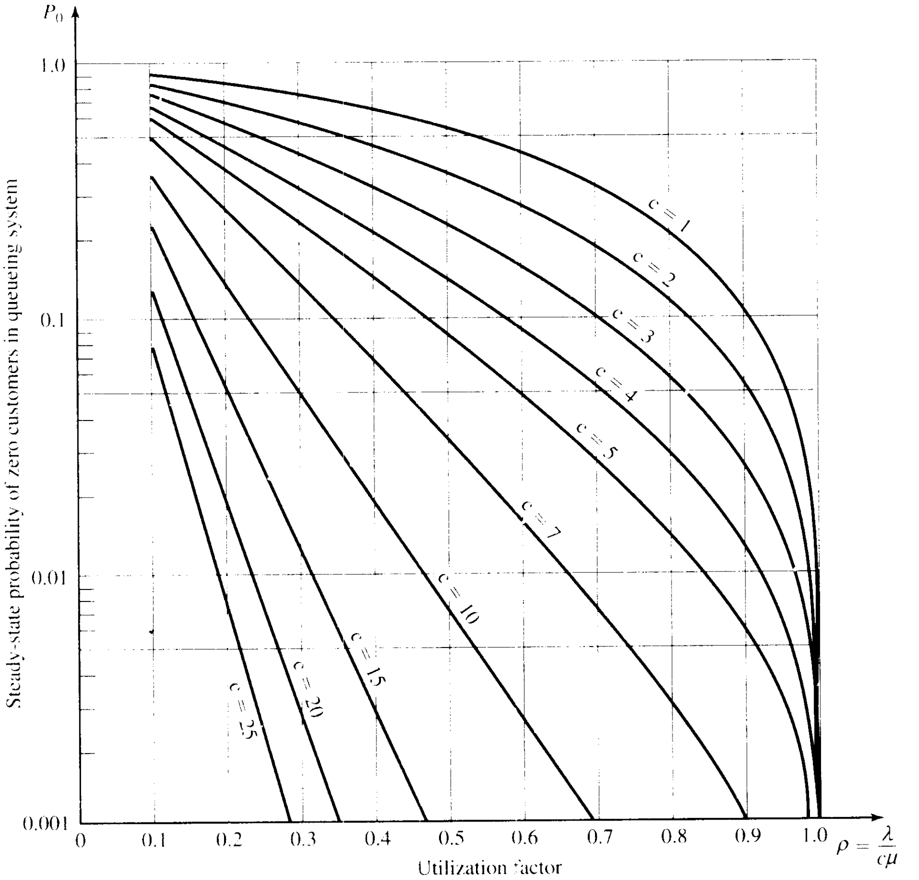

Figure 6.14 Values of P_0 for $M/M/c/\infty$ model. (From F S. Hillier and G. J. Lieberman, *Introduction to Operations Research*, 5th ed., 1990, p. 616. Adapted with permission of McGraw-Hill, Inc., New York.)

Thus, the time-average length of the waiting line of mechanics is

$$L_Q = \frac{(2/3)(8/15)}{1-2/3} = 1.07 \text{ mechanics}$$

and the time-average number in system is given by

$$L = L_Q + \frac{\lambda}{\mu} = \frac{16}{15} + \frac{4}{3} = \frac{12}{5} = 2.4 \text{ mechanics}$$

From Little's relationships, the average time a mechanic spends at the tool crib is

$$w = \frac{L}{\lambda} = \frac{2.4}{2} = 1.2 \text{ minutes}$$

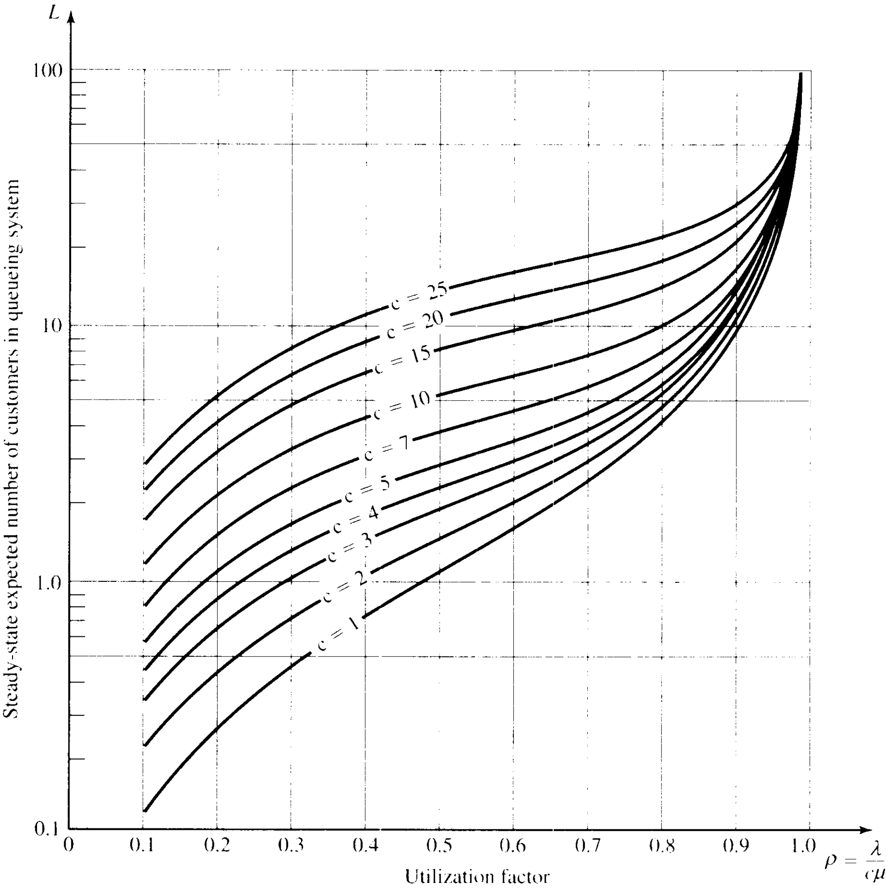

Figure 6.15 Values of L for $M/M/c/\infty$ model. (From F. S. Hillier and G. J. Lieberman, *Introduction to Operations Research*, 5th ed., 1990, p. 617. Adapted with permission of McGraw-Hill, Inc., New York.)

and the average time spent waiting for an attendant is

$$w_Q = w - \frac{1}{\mu} = 1.2 - \frac{2}{3} = 0.533 \text{ minute}$$

Example 6.14

Using the data of Example 6.13, compute P_0 and L from Figures 6.14 and 6.15. First, compute

$$\rho = \frac{\lambda}{c\mu} = \frac{2}{2(3/2)} = \frac{2}{3} = 0.667$$

Entering the utilization factor 0.667 on the horizontal axis of Figure 6.14 gives the value 0.2 for P_0 on the vertical axis. Similarly, the value $L = 2.4$ is read from the vertical axis of Figure 6.15.

An Approximation for the M/G/c/∞ Queue

Recall that formulas for L_Q and w_Q for the *M/G/*1 queue can be obtained from the corresponding *M/M/*1 formulas by multiplying them by the correction factor $(1 + (\text{cv})^2)/2$, as in Equation (6.18). *Approximate* formulas for the *M/G/c* queue can be obtained by applying the same correction factor to the *M/M/c* formulas for L_Q and w_Q (no exact formula exists for $1 < c < \infty$). The nearer the cv is to 1, the better the approximation.

Example 6.15

Recall Example 6.13. Suppose that the service times for the mechanics at the tool crib are not exponentially distributed, but are known to have a standard deviation of 30 seconds. Then we have an *M/G/c* model, rather than an *M/M/c*. The mean service time is 40 seconds, so the coefficient of variation of the service time is

$$\text{cv} = \frac{30}{40} = \frac{3}{4} < 1$$

Therefore, the accuracy of L_Q and w_Q can be improved by the correction factor

$$\frac{1+(\text{cv})^2}{2} = \frac{1+(3/4)^2}{2} = \frac{25}{32} = 0.78$$

For example, when there are $c = 2$ attendants,

$$L_Q = (0.78)(1.07) = 0.83 \text{ mechanics}$$

Notice that, because the coefficient of variation of the service time is less than 1, the congestion in the system, as measured by L_Q, is less than in the corresponding *M/M/*2 model.

The correction factor applies only to the formulas for L_Q and w_Q. Little's formula can then be used to calculate L and w. Unfortunately, there is no general method for correcting the steady-state probabilities, P_n.

When the Number of Servers is Infinite (M/G/∞/∞)

There are at least three situations in which it is appropriate to treat the number of servers as infinite:

1. when each customer is its own server—in other words, in a self-service system;
2. when service capacity far exceeds service demand, as in a so-called ample-server system; and
3. when we want to know how many servers are required so that customers will rarely be delayed.

The steady-state parameters for the *M/G/*∞ queue are listed in Table 6.6. In the table, λ is the arrival rate of the Poisson arrival process, and $1/\mu$ is the expected service time of the general service-time distribution (including exponential, constant, or any other).

Example 6.16

Prior to introducing their new, subscriber-only, on-line computer information service, The Connection must plan their system capacity in terms of the number of users that can be logged in simultaneously. If the service is successful, customers are expected to log on at a rate of $\lambda = 500$ per hour, according to a Poisson process, and stay connected for an average of $1/\mu = 180$ minutes (or 3 hours). In the real system, there will be an upper limit on simultaneous users, but, for planning purposes, The Connection can pretend that the number of simultaneous users is infinite. An *M/G/*∞ model of the system implies that the expected number of simultaneous users is $L = \lambda/\mu = 500(3) = 1500$, so a capacity greater than 1500 is certainly required. To ensure providing adequate capacity 95% of the time, The Connection could allow the number of simultaneous users to be the smallest value c such that

Table 6.6 Steady-State Parameters for the $M/G/\infty$ Queue

P_0	$e^{-\lambda/\mu}$
w	$\frac{1}{\mu}$
w_Q	0
L	$\frac{\lambda}{\mu}$
L_Q	0
P_n	$\frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}, \ n = 0, 1, \ldots$

$$P(L(\infty) \le c) = \sum_{n=0}^{c} P_n = \sum_{n=0}^{c} \frac{e^{-1500}(1500)^n}{n!} \ge 0.95$$

The capacity $c = 1564$ simultaneous users satisfies this requirement.

6.4.3 Multiserver Queues with Poisson Arrivals and Limited Capacity: M/M/c/N/∞

Suppose that service times are exponentially distributed at rate μ, that there are c servers, and that the total system capacity is $N \ge c$ customers. If an arrival occurs when the system is full, that arrival is turned away and does not enter the system. As in the preceding section, suppose that arrivals occur randomly according to a Poisson process with rate λ arrivals per time unit. For any values of λ and μ such that $\rho \ne 1$, the $M/M/c/N$ queue has a statistical equilibrium with steady-state characteristics as given in Table 6.7 (formulas for the case $\rho = 1$ can be found in Hillier and Lieberman [2005]).

Table 6.7 Steady-State Parameters for the $M/M/c/N$ Queue (N = System Capacity, $a = \lambda/\mu$, $\rho = \lambda/(c\mu)$)

P_0	$\left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^{N} \rho^{n-c}\right]^{-1}$
P_N	$\frac{a^N}{c!c^{N-c}} P_0$
L_Q	$\frac{P_0 a^c \rho}{c!(1-\rho)^2}\left[1-\rho^{N-c}-(N-c)\rho^{N-c}(1-\rho)\right]$
λ_e	$\lambda(1-P_N)$
w_Q	$\frac{L_Q}{\lambda_e}$
w	$w_Q + \frac{1}{\mu}$
L	$\lambda_e w$

The effective arrival rate, λ_e, is defined as the mean number of arrivals per time unit who enter and remain in the system. For all systems, $\lambda_e \le \lambda$; for the unlimited-capacity systems, $\lambda_e = \lambda$; but, for systems such as the present one, which turn customers away when full, $\lambda_e < \lambda$. The effective arrival rate is computed by

$$\lambda_e = \lambda\,(1 - P_N)$$

because $1 - P_N$ is the probability that a customer, upon arrival, will find space and be able to enter the system. When one is using Little's equations (6.17) to compute mean time spent in system w and in queue w_Q, λ must be replaced by λ_e.

Example 6.17

The unisex hair-styling shop described in Example 6.17 can hold only three customers: one in service, and two waiting. Additional customers are turned away when the system is full. The offered load is as previously determined, namely $\lambda/\mu = 2/3$.

In order to calculate the performance measures, first compute P_0:

$$P_0 = \left[1 + \frac{2}{3} + \frac{2}{3}\sum_{n=2}^{3}\left(\frac{2}{3}\right)^{n-1}\right]^{-1} = 0.415$$

The probability that there are three customers in the system (the system is full) is

$$P_N = P_3 = \frac{(2/3)^3}{1!1^2}P_0 = \frac{8}{65} = 0.123$$

Then, the average length of the queue (customers waiting for a haircut) is given by

$$L_Q = \frac{(27/65)(2/3)(2/3)}{(1-2/3)^2}[1-(2/3)^2 - 2(2/3)^2(1-2/3)] = 0.431 \text{ customer}$$

Now, the effective arrival rate, λ_e, is given by

$$\lambda_e = 2\left(1 - \frac{8}{65}\right) = \frac{114}{65} = 1.754 \text{ customers per hour}$$

Therefore, from Little's equation, the expected time spent waiting in queue is

$$w_Q = \frac{L_Q}{\lambda_e} = \frac{28}{114} = 0.246 \text{ hour}$$

and the expected total time in the shop is

$$w = w_Q + \frac{1}{\mu} = \frac{66}{114} = 0.579 \text{ hour}$$

One last application of Little's equation gives the expected number of customers in the shop (in queue and getting a haircut) as

$$L = \lambda_e w = \frac{66}{65} = 1.015 \text{ customers}$$

Notice that $1 - P_0 = 0.585$ is the average number of customers being served or, equivalently, the probability that the single server is busy. Thus, the server utilization, or proportion of time the server is busy in the long run, is given by

$$1 - P_0 = \frac{\lambda_e}{\mu} = 0.585$$

The reader should compare these results to those of the unisex hair-styling shop before the capacity constraint was placed on the system. Specifically, in systems with limited capacity, the offered load λ/μ can assume any positive value and no longer equals the server utilization $\rho = \lambda_e/\mu$. Notice that server utilization decreases from 67% to 58.5% when the system imposes a capacity constraint.

6.5 STEADY-STATE BEHAVIOR OF FINITE-POPULATION MODELS (M/M/C/K/K)

In many practical problems, the assumption of an infinite calling population leads to invalid results because the calling population is, in fact, small. When the calling population is small, the presence of one or more customers in the system has a strong effect on the distribution of future arrivals, and the use of an infinite-population model can be misleading. Typical examples include a small group of machines that break down from time to time and require repair, or a small group of mechanics who line up at a counter for parts or tools. In the extreme case, if all the machines are broken, no new "arrivals" (breakdowns) of machines can occur; similarly, if all the mechanics are in line, no arrival is possible to the tool and parts counter. Contrast this to the infinite-population models, in which the arrival rate, λ, of customers to the system is assumed to be independent of the state of the system.

Consider a finite-calling-population model with K customers. The time between the end of one service visit and the next call for service for each member of the population is assumed to be exponentially distributed, with mean $1/\lambda$ time units; service times are also exponentially distributed, with mean $1/\mu$ time units; there are c parallel servers, and system capacity is K, so that all arrivals remain for service. Such a system is depicted in Figure 6.16.

The steady-state parameters for this model are listed in Table 6.8. An electronic spreadsheet or a symbolic calculation program is useful for evaluating these complex formulas. For example, Figure 6.17 is a procedure written for the symbolic calculation program Maple to calculate the steady-state probabilities for the *M/M/c/K/K* queue. Another approach is to use precomputed queueing tables, such as those found in Banks and Heikes [1984], Hillier and Yu [1981], Peck and Hazelwood [1958] or Descloux [1962].

The effective arrival rate λ_e has several valid interpretations:

λ_e = long-run effective arrival rate of customers to the queue
= long-run effective arrival rate of customers entering service
= long-run rate at which customers exit from service
= long-run rate at which customers enter the calling population
(and begin a new runtime)
= long-run rate at which customers exist from the calling population

Example 6.18

There are two workers who are responsible for 10 milling machines. The machines run on the average for 20 minutes, then require an average 5-minute service period, both times exponentially distributed. Therefore, $\lambda = 1/20$ and $\mu = 1/5$. Compute the various measures of performance for this system.

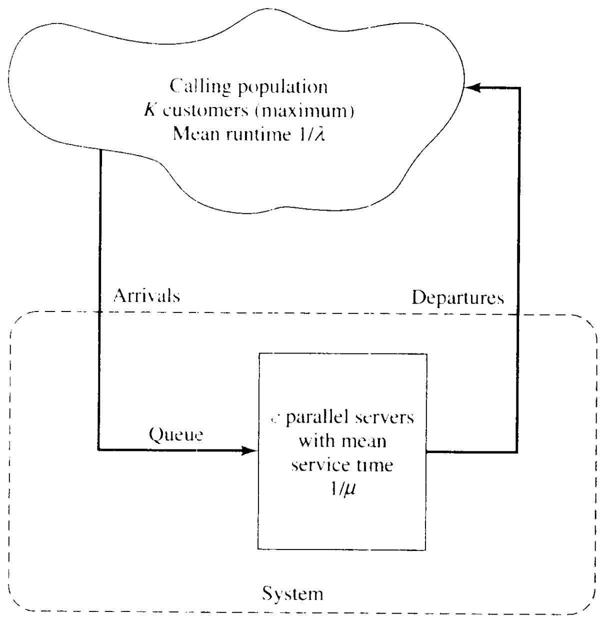

Figure 6.16 Finite-population queueing model.

Table 6.8 Steady-State Parameters for the M/M/*c*/*K*/*K* Queue

P_0	$\left[\sum_{n=0}^{c-1}\binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n+\sum_{n=c}^{K}\frac{K!}{(K-n)!c!c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$
P_n	$\begin{cases}\binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n P_0, & n=0,1,\ldots,c-1\\ \frac{K!}{(K-n)!c!c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n P_0, & n=c,c+1,\ldots,K\end{cases}$
L	$\sum_{n=0}^{K} nP_n$
L_Q	$\sum_{n=c+1}^{K}(n-c)P_n$
λ_e	$\sum_{n=0}^{K}(K-n)\lambda P_n$
w	L/λ_e
w_Q	L_Q/λ_e
ρ	$\frac{L-L_Q}{c}=\frac{\lambda_e}{c\mu}$

```
mmcKK := proc(lambda, mu, c, K)
          # return steady-state probabilities for M/M/c/K/K queue
          # notice that p[n+1] is P_n, n=0,..,K
          local crho, Kfac, cfac, p, n;
          p : = vector(K+1,0);
          crho := lambda/mu;
          Kfac := K!;
          cfac := c!;
          p[1] := sum((Kfac/(n!*(K-n)!))*crho^n, n=0..c-1) + sum((Kfac/(c^(n-c)*
                (K-n)!*cfac))*crho^n, n=c..K);
          p[1] := 1/p[1];
          for n from 1 to c-1
          do
             p[n+1] := p[1]*(Kfac/(n!*(K-n)!))*crho^n;
          od;
          for n from c to K
          do
             p[n+1] := p[1]*(Kfac/(c^(n-c)*(K-n)!*cfac))*crho^n;
          od;
          RETURN(evalm(p));
          end;
```

Figure 6.17 Maple procedure to calculate P_n for the $M/M/c/K/K$ queue.

All of the performance measures depend on P_0, which is

$$\left[\sum_{n=0}^{2-1}\binom{10}{n}\left(\frac{5}{20}\right)^n+\sum_{n=c}^{10}\frac{10!}{(10-n)!2!2^{n-2}}\left(\frac{5}{20}\right)^n\right]^{-1}=0.065$$

From P_0, we can obtain the other P_n, from which we can compute the average number of machines waiting for service,

$$L_Q=\sum_{n=3}^{10}(n-2)P_n=1.46 \text{ machines}$$

the effective arrival rate,

$$\lambda_e=\sum_{n=0}^{10}(10-n)\left(\frac{1}{20}\right)P_n=0.342 \text{ machines/minute}$$

and the average waiting time in the queue,

$$w_Q=L_Q/\lambda_e=4.27 \text{ minutes}$$

Similarly, we can compute the expected number of machines being serviced or waiting to be serviced,

$$L=\sum_{n=0}^{10}nP_n=3.17 \text{ machines}$$

The average number of machines being serviced is given by

$$L-L_Q=3.17-1.46=1.71 \text{ machines}$$

Each machine must be either running, waiting to be serviced, or in service, so the average number of running machines is given by

$$K - L = 10 - 3.17 = 6.83 \text{ machines}$$

A question frequently asked is this: What will happen if the number of servers is increased or decreased? If the number of workers in this example increases to three ($c = 3$), then the time-average number of running machines increases to

$$K - L = 7.74 \text{ machines}$$

an increase of 0.91 machine, on the average.

Conversely, what happens if the number of servers decreases to one? Then the time-average number of running machines decreases to

$$K - L = 3.98 \text{ machines}$$

The decrease from two servers to one has resulted in a drop of nearly three machines running, on the average. Examples 12, and 20 asks the reader to determine the optimal number of servers.

Example 6.18 illustrates several general relationships that have been found to hold for almost all queues. If the number of servers is decreased, delays, server utilization, and the probability of an arrival having to wait to begin service all increase.

6.6 NETWORKS OF QUEUES

In this chapter, we have emphasized the study of single queues of the *G/G/c/N/K* type. However, many systems are naturally modeled as networks of single queues in which customers departing from one queue may be routed to another. Example 6.1 (see, in particular, Figure 6.3) and Example 6.2 (see Figure 6.5) are illustrations.

The study of mathematical models of networks of queues is beyond the scope of this chapter; see, for instance, Gross and Harris [1997], Nelson [1995], and Kleinrock [1976]. However, a few fundamental principles are very useful for rough-cut modeling, perhaps prior to a simulation study. The following results assume a stable system with infinite calling population and no limit on system capacity:

1. Provided that no customers are created or destroyed in the queue, then the departure rate out of a queue is the same as the arrival rate into the queue, over the long run.
2. If customers arrive to queue i at rate λ_i, and a fraction $0 \le p_{ij} \le 1$ of them are routed to queue j upon departure, then the arrival rate from queue i to queue j is $\lambda_i p_{ij}$ over the long run.
3. The overall arrival rate into queue j, λ_j, is the sum of the arrival rate from all sources. If customers arrive from outside the network at rate a_j, then

$$\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$$

4. If queue j has $c_j < \infty$ parallel servers, each working at rate μ_j, then the long-run utilization of each server is

$$\rho_j = \frac{\lambda_j}{c_j \mu_j}$$

and $\rho_j < 1$ is required for the queue to be stable.

5. If, for each queue j, arrivals from outside the network form a Poisson process with rate a_j, and if there are c_j identical servicers delivering exponentially distributed service times with mean $1/\mu_j$ (where c_j may be ∞), then, in steady state, queue j behaves like an $M/M/c_j$ queue with arrival rate $\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$.

Example 6.19

Consider again the discount store described in Example 6.1 and shown in Figure 6.3. Suppose that customers arrive at the rate 80 per hour and that, of those arrivals, 40% choose self-service; then, the arrival rate to service center 1 is $\lambda_1 = (80)(0.40) = 32$ per hour, and the arrival rate to service center 2 is $\lambda_2 = (80)(0.6) = 48$ per hour. Suppose that each of the $c_2 = 3$ clerks at service center 2 works at the rate $\mu_2 = 20$ customers per hour. Then the long-run utilization of the clerks is

$$\rho_2 = \frac{48}{(3)(20)} = 0.8$$

All customers must see the cashier at service center 3. The overall arrival rate to service center 3 is $\lambda_3 = \lambda_1 + \lambda_2 = 80$ per hour, regardless of the service rate at service center 1, because, over the long run, the departure rate out of each service center must be equal to the arrival rate into it. If the cashier works at rate $\mu_3 = 90$ per hour, then the utilization of the cashier is

$$\rho_3 = \frac{80}{90} = 0.89$$

Example 6.20

At a Driver's License branch office, drivers arrive at the rate 50 per hour. All arrivals must first check in with one of two clerks, with the average check-in time being 2 minutes. After check in, 15% of the drivers need to take a written test that lasts approximately 20 minutes. All arrivals must wait to have their picture taken and their license produced; this station can process about 60 drivers per hour. The branch manager wants to know whether it is adding a check-in clerk or adding a new photo station that will lead to a greater reduction in customer delay.

To solve the problem, let the check-in clerks be queue 1 (with $c_1 = 2$ servers, each working at rate $\mu_1 = 30$ drivers per hour), let the testing station be queue 2 (with $c_2 = \infty$ servers, because any number of people can be taking the written test simultaneously, and service rate $\mu_2 = 3$ drivers per hour), and let the photo station be queue 3 (with $c_3 = 1$ server working at rate $\mu_3 = 60$ drivers per hour). The arrival rates to each queue are as follows:

$$\lambda_1 = a_1 + \sum_{i=1}^{3} p_{i1}\lambda_i = 50 \text{ drivers per hour}$$

$$\lambda_2 = a_2 + \sum_{i=1}^{3} p_{i2}\lambda_i = (0.15)\lambda_1 \text{ drivers per hour}$$

$$\lambda_3 = a_3 + \sum_{i=1}^{3} p_{i3}\lambda_i = (1)\lambda_2 + (0.85)\lambda_1 \text{ drivers per hour}$$

Notice that arrivals from outside the network occur only at queue 1, so $a_1 = 50$ and $a_2 = a_3 = 0$. Solving this system of equations gives $\lambda_1 = \lambda_3 = 50$ and $\lambda_2 = 7.5$.

If we approximate the arrival process as Poisson, and the service times at each queue as exponentially distributed, then the check-in clerks can be approximated as an $M/M/c_1$ queue, the testing station as an $M/M/\infty$

queue, and the photo station as an $M/M/c_3$ queue. Thus, under the current set-up, the check-in station is an $M/M/2$; using the formulas in Table 6.5 gives $w_Q = 0.0758$ hours. If we add a clerk, so that the model is $M/M/3$, the waiting time in queue drops to 0.0075 hours, a savings of 0.0683 hours or about 4.1 minutes.

The current photo station can be modeled as an $M/M/1$ queue, giving $w_Q = 0.0833$ hours; adding a second photo station ($M/M/2$) causes the time in queue to drop to 0.0035 hours, a savings of 0.0798 hours, or about 4.8 minutes. Therefore, a second photo station offers a slightly greater reduction in waiting time than does adding a third clerk.

If desired, the testing station can be analyzed by using the results for an $M/M/\infty$ queue in Table 6.6. For instance, the expected number of people taking the test at any time is $L = \lambda_2/\mu_2 = 7.5/3 = 2.5$.

6.7 SUMMARY

Queueing models have found widespread use in the analysis of service facilities, production and material-handling systems, telephone and communications systems, and many other situations where congestion or competition for scarce resources can occur. This chapter has introduced the basic concepts of queueing models and shown how simulation, and in some cases a mathematical analysis, can be used to estimate the performance measures of a system.

A simulation may be used to generate one or more artificial histories of a complex system. This simulation-generated data may, in turn, be used to estimate desired performance measures of the system. Commonly used performance measures, including L, L_Q, w, w_Q, ρ, and λ_e were introduced, and formulas were given for their estimation from data.

When simulating any system that evolves over time, the analyst must decide whether transient behavior or steady-state performance is to be studied. Simple formulas exist for the steady-state behavior of some queues, but estimating steady-state performance measures from simulation-generated data requires recognizing and dealing with the possibly deleterious effect of the initial conditions on the estimators of steady-state performance. These estimators could be severely biased (either high or low), if the initial conditions are unrepresentative of steady state or if simulation run length is too short. These estimation problems are discussed at greater length in Chapter 11.

Whether the analyst is interested in transient or in steady-state performance of a system, it should be recognized that the estimates obtained from a simulation of a stochastic queue are exactly that—estimates. Every such estimate contains random error, and a proper statistical analysis is required to assess the accuracy of the estimate. Methods for conducting such a statistical analysis are discussed in Chapters 11 and 12.

In the last three sections of this chapter, it was shown that a number of simple models can be solved mathematically. Although the assumptions behind such models might not be met exactly in a practical application, these models can still be useful in providing a rough estimate of a performance measure. In many cases, models with exponentially distributed interarrival and service times will provide a conservative estimate of system behavior. For example, if the model predicts that average waiting time, w, will be 12.7 minutes, then average waiting time in the real system is likely to be less than 12.7 minutes. The conservative nature of exponential models arises because (a) performance measures, such as w and L, are generally increasing functions of the variance of interarrival times and service times (recall the $M/G/1$ queue), and (b) the exponential distribution is fairly highly variable, having its standard deviation always equal to its mean. Thus, if the arrival process or service mechanism of the real system is less variable than exponentially distributed interarrival or service times, it is likely that the average number in the system, L, and the average time spent in system, w, will be less than what is predicted by the exponential model. Of course, if the interarrival and service times are *more* variable than exponential random variables, then the M/M queueing models could underestimate congestion.

An important application of mathematical queueing models is determining the minimum number of servers needed at a work station or service center. Quite often, if the arrival rate λ and the service rate μ are known or can be estimated, then the simple inequality $\lambda/(c\mu) < 1$ can be used to provide an initial estimate for the number of servers, c, at a work station. For a large system with many work stations, it could be quite time consuming to have to simulate every possibility $(c_1, c_2, \ldots)$ for the number of servers, c_i, at work station i. Thus, a bit of mathematical analysis rough estimates could save a great deal of computer time and analyst's time.

Finally, the qualitative behavior of the simple exponential models of queueing carries over to more complex systems. In general, it is the variability of service times and the variability of the arrival process that causes waiting lines to build up and congestion to occur. For most systems, if the arrival rate increases, or if the service rate decreases, or if the variance of service times or interarrival times increases, then the system will become more congested. Congestion can be decreased by adding more servers or by reducing the mean value and variability of service times. Simple queueing models can be a great aid in quantifying these relationships and in evaluating alternative system designs.

REFERENCES

BANKS, J., AND R. G. HEIKES [1984], *Handbook of Tables and Graphs for the Industrial Engineer and Manager*, Reston, Reston, VA.

COOPER, R. B. [1990], *Introduction to Queueing Theory*, 3d ed., George Washington University, Washington, DC.

DESCLOUX, A. [1962], *Delay Tables for Finite- and Infinite-Source Systems*, McGraw–Hill, New York.

GROSS, D., AND C. HARRIS [1997], *Fundamentals of Queueing Theory*, 3d ed., Wiley, New York.

HALL, R. W. [1991], *Queueing Methods: For Services and Manufacturing*, Prentice Hall, Englewood Cliffs, NJ.

HILLIER, F. S., AND G. J. LIEBERMAN [2005], *Introduction to Operations Research*, 8th ed., McGraw-Hill New York.

HILLIER, F. S., AND O. S. YU [1981], *Queueing Tables and Graphs*, Elsevier North-Holland, New York.

KENDALL, D. G. [1953], "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains," *Annals of Mathematical Statistics*, Vol. 24, pp. 338–354.

KLEINROCK, L. [1976], *Queueing Systems, Vol 2: Computer Applications*, Wiley, New York.

LITTLE, J. D. C. [1961], "A Proof for the Queueing Formula $L = \lambda w$," *Operations Research*, Vol. 16, pp. 651–665.

NELSON, B. L. [1995], *Stochastic Modeling: Analysis & Simulation*, Dover Publications, Mineola, NY.

PECK, L. G., AND R. N. HAZELWOOD [1958], *Finite Queueing Tables*, Wiley, New York.

WAGNER, H. M. [1975], *Principles of Operations Research*, 2d ed., Prentice-Hall, Englewood Cliffs, NJ.

WINSTON, W. L. [1997], *Operations Research: Applications and Algorithms*, Duxbury Press, Pacific Grove, CA.

EXERCISES

1. Identify the calling population, customer, and server in the following queueing situations:

- **(a)** university library
- **(b)** bank teller counter
- **(c)** Internet router
- **(d)** police station
- **(e)** assembly line

2. A two-runway (one runway for landing, one runway for taking off) airport is being designed for propeller-driven aircraft. The time to land an airplane is known to be exponentially distributed, with a mean of 1-1/2 minutes. If airplane arrivals are assumed to occur at random, what arrival rate can be tolerated if the average wait in the sky is not to exceed 3 minutes?

3. If customers arrive for service according to Poisson distribution with a mean of 5 per day, how fast the average service time (assume exponential) must be to keep average number in the system less than 4?

4. Give some examples from real-life situations for balking and reneging.

5. Trucks arrive at a facility to be unloaded in a pattern, which can be characterized by the Poisson distribution. The average rate of arrivals is 36 per hour, and the level of service is exponentially distributed with a mean service rate of 39 trucks per hour. Compute all the relevant statistics for the system. The drivers make Rs. 9 each hour and do not unload the trucks. How much expense, on the average, is incurred by the trucking company for idle time on the part of each driver for each visit to the facility?

6. Patients arrive for a physical examination according to a Poisson process at the rate 1 per hour. The physical examination requires three stages, each one independently exponentially distributed, with a service time of 15 minutes. A patient must go through all three stages before the next patient is admitted to the treatment facility. Compute the average number of delayed patients, L_Q, for this system. (Hint: The variance of the sum of independent random variables is the sum of the variance.)

7. Suppose that mechanics arrive randomly at a tool crib according to a Poisson process with rate $\lambda = 10$ per hour. It is known that the single tool clerk serves a mechanic in 4 minutes on the average, with a standard deviation of approximately 2 minutes. Suppose that mechanics make $15.00 per hour. Estimate the steady-state average cost per hour of mechanics waiting for tools.

8. The arrival of customers at a teller counter follows Poisson with a mean of 45 per hour and teller's service time follows exponential with a mean of 1 minute. Determine the following:

(a) Probability of having 0 customer in the system, 5 customers in the system, and 10 customers in the system.

(b) Determine L_Q, L, W_Q, and W.

9. A machine shop repairs small electric motors which arrive according to a Poisson process at the rate 12 per week (5-day, 40-hour workweek). An analysis of past data indicates that engines can be repaired, on the average, in 2.5 hours, with a variance of 1 hour^2. How many working hours should a customer expect to leave a motor at the repair shop (not knowing the status of the system)? If the variance of the repair time could be controlled, what variance would reduce the expected waiting time to 6.5 hours?

10. Arrivals to a self-service gasoline pump occur in a Poisson fashion at the rate 12 per hour. Service time has a distribution that averages 4 minutes, with a standard deviation of 1-1/3 minutes. What is the expected number of vehicles in the system?

11. Classic Car Care has one worker who washes cars in a four-step method—soap, rinse, dry, vacuum. The time to complete each step is exponentially distributed, with mean 9 minutes. Every car goes through every step before another car begins the process. On the average, one car every 45 minutes arrives for a wash job, according to a Poisson process. What is the average time a car waits to begin the wash job? What is the average number of cars in the car wash system? What is the average time required to wash a car?

12. Machines arrive for repair at the rate of six per hour following Poisson. The mechanics mean repair time is 15 minutes, which follows exponential distribution. The down time cost for the broken down machines per hour is Rs. 300. Mechanics are paid Rs. 60 per hour. Determine the optimal number of mechanics to be employed to minimize the total cost.

13. Given the following information for a finite calling population problem with exponentially distributed runtimes and service times:

$$K = 10$$
$$\frac{1}{\mu} = 15$$
$$\frac{1}{\lambda} = 82$$
$$c = 2$$

Compute L_Q and w_Q. Find the value of λ such that $L_Q = L/2$.

14. Suppose that Figure 6.6 represents the number in system for a last-in–first-out (LIFO) single-server system. Customers are not preempted (i.e., kicked out of service), but, upon service completion, the most recent arrival next begins service. For this LIFO system, apportion the total area under $L(t)$ to each individual customer, as was done in Figure 6.8 for the FIFO system. Using the figure, show that Equations (6.10) and (6.8) hold for the single-server LIFO system.

15. Repeat Exercise 14, but assuming that

(a) Figure 6.6 represents a FIFO system with $c = 2$ servers:
(b) Figure 6.6 represents a LIFO system with $c = 2$ servers:

16. Consider a $M/G/1$ queue with the following type of service distribution: Customers request one of two types of service, in the proportions p and $1-p$. Type i service is exponentially distributed at rate μ_i, $i = 1, 2$. Let X_i denote a type-i service time and X an arbitrary service time. Then $E(X_i) = 1/\mu_i$, $V(X_i) = 1/\mu_i^2$ and

$$X = \begin{cases} X_1 & \text{with probability } p \\ X_2 & \text{with probability } (1-p) \end{cases}$$

The random variable X is said to have a hyperexponential distribution with parameters (μ_1, μ_2, p).

(a) Show that $E(X) = p/\mu_1 + (1-p)/\mu_2$ and $E(X^2) = 2p/\mu_1^2 + 2(1-p)/\mu_2^2$.
(b) Use $V(X) = E(X^2) - [E(X)]^2$ to show $V(X) = 2p/\mu_1^2 + 2(1-p)/\mu_2^2 - [p/\mu_1 + (1-p)/\mu_2]^2$.
(c) For any hyperexponential random variable, if $\mu_1 \neq \mu_2$ and $0 < p < 1$, show that its coefficient of variation is greater than 1—that is, $(\text{cv})^2 = V(X)/[E(X)]^2 > 1$. Thus, the hyperexponential distribution provides a family of statistical models for service times that are more variable than exponentially distributed service times. *Hint*: The algebraic expression for $(\text{cv})^2$, by using parts (a) and (b), can be manipulated into the form $(\text{cv})^2 = 2p(1-p)(1/\mu_1 - 1/\mu_2)^2/[E(X)]^2 + 1$.
(d) Many choices of μ_1, μ_2, and p lead to the same overall mean $E(X)$ and $(\text{cv})^2$. If a distribution with mean $E(X) = 1$ and coefficient of variation cv $= 2$ is desired, find values of μ_1, μ_2, and p to achieve this. *Hint*: Choose $p = 1/4$ arbitrarily; then solve the following equations for μ_1 and μ_2.

$$\frac{1}{4\mu_1} + \frac{3}{4\mu_2} = 1$$

$$\frac{3}{8}\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)^2 + 1 = 4$$

17. Orders are expected to arrive at a machining center according to Poisson process at a mean rate of 30 per hour. The management has an option of two machines M1 (fast but expensive) and M2 (slow inexpensive). Both machines would have an exponential distribution for machining times with M1 having a mean of 1.2 minutes and M2 having a mean of 1.5 minutes. The profit per year is given by Rs. 72,000/W, where W is the expected waiting time (in minutes) for the orders in the system. Determine the upper bound on the difference in the average yearly cost that would justify buying M1 rather than M2.

18. In Example 6.18, increase the number of machines by 2, then compare the systems with $c = 1$, $c = 2$, and $c = 3$ servers on the basis of server utilization ρ (the proportion of time a typical server is busy).

19. Vehicles pass through a toll gate at a rate of 90 per hour. The average time to pass through the gate is 36 seconds. The arrival rate and service rate follow Poisson distribution. There is a complaint that the vehicles wait for a long duration. The authorities are willing to install one more gate to reduce the average time to pass through to 30 seconds, if the idle time of the toll gate is less than 10% and the present average queue length at the gate is more than five vehicles. Check whether the installation of the second gate is justified.

20. The arrival of employees at a tool crib can be described by a Poisson distribution. Service times are exponentially distributed. The rate of arrival averages 45 machinists per hour, while an attendant can serve an average of 50 men per hour. The machinists are paid Rs. 24 per hour, while the attendants are paid Rs. 15 per hour. Find the optimum number of attendants to place in the crib, assuming 8 hours and 200 days per year.

21. This problem is based on Case 8.1 in Nelson [1995]. A large consumer shopping mall is to be constructed. During busy times, the arrival rate of cars is expected to be 1000 per hour, and studies at other malls suggest that customers will spend 3 hours, on average, shopping. The mall designers would like to have sufficient parking so that there are enough spaces 99.9% of the time. How many spaces should they have? Hint: Model the system as an M/G/∞ queue where the spaces are servers, and find out how many spaces are adequate with probability 0.999.

22. In Example 6.19, suppose that the overall arrival rate is expected to increase to 160 per hour. If the service rates do not change, how many clerks will be needed at service centers 2 and 3, just to keep up with the customer load?

23. A small copy shop has a self-service copier. Currently there is room for only 4 people to line up for the machine (including the person using the machine); when there are more than 4 people, then the additional people must line up outside the shop. The owners would like to avoid having people line up outside the shop, as much as possible. For that reason, they are thinking about adding a second self-service copier. Self-service customers have been observed to arrive at the rate 24 per hour, and they use the machine 2 minutes, on average. Assess the impact of adding another copier. Carefully state any assumptions or approximations you make.

24. In an N machine one operator environment, five automatic machines are attended by one operator. Every time a machine completes a batch, the operator must reset it before a new batch is started. The time to complete a batch run is exponential with a mean of 45 minutes. The setup time is also exponential with a mean of 8 minutes. Determine

(a) the average number of machines that are waiting for set up.
(b) the probability that all the machines are working.
(c) the average time a machine is down.

25. Search the web and find applications of queueing theory in production activities.

26. Study the effect of *pooling servers* (having multiple servers draw from a single queue, rather than each having its own queue) by comparing the performance measures for two M/M/1 queues, each with arrival rate λ and service rate μ, to an M/M/2 queue with arrival rate 2λ and service rate μ for each server.

27. A repair and inspection facility consists of two stations: a repair station with two technicians, and an inspection station with 1 inspector. Each repair technician works at the rate 3 items per hour; the inspector can inspect 8 items per hour. Approximately 10% of all items fail inspection and are sent back to the repair station. (This percentage holds even for items that have been repaired two or more times.) If items arrive at the rate 5 per hour, what is the long-run expected delay that items experience at each of the two stations, assuming a Poisson arrival process and exponentially distributed service times? What is the maximum arrival rate that the system can handle without adding personnel?

Part III

Random Numbers

7

Random-Number Generation

Random numbers are a necessary basic ingredient in the simulation of almost all discrete systems. Most computer languages have a subroutine, object, or function that will generate a random number. Similarly, simulation languages generate random numbers that are used to generate event times and other random variables. In this chapter, the generation of random numbers and their subsequent testing for randomness is described. Chapter 8 shows how random numbers are used to generate a random variable with any desired probability distribution.

7.1 PROPERTIES OF RANDOM NUMBERS

A sequence of random numbers, R_1, R_2, ..., must have two important statistical properties: uniformity and independence. Each random number R_i must be an independent sample drawn from a continuous uniform distribution between zero and 1—that is, the pdf is given by

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

This density function is shown in Figure 7.1. The expected value of each R_i is given by

$$E(R) = \int_0^1 x\,dx = \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2}$$

and the variance is given by

$$V(R) = \int_0^1 x^2\,dx - \left[E(R)\right]^2 = \left.\frac{x^3}{3}\right|_0^1 - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Usually, random numbers are generated by a digital computer, as part of the simulation. There are numerous methods that can be used to generate the values. Before we describe some of these methods, or *routines*, there are a number of important considerations that we should mention:

1. The routine should be fast. Individual computations are inexpensive, but simulation could require many millions of random numbers. The total cost can be managed by selecting a computationally efficient method of random-number generation.
2. The routine should be portable to different computers—and, ideally, to different programming languages. This is desirable so that the simulation program will produce the same results wherever it is executed.
3. The routine should have a sufficiently long cycle. The cycle length, or period, represents the length of the random number sequence before previous numbers begin to repeat themselves in an earlier order. Thus, if 10,000 events are to be generated, the period should be many times that long.
   A special case of cycling is degenerating. A routine degenerates when the same random numbers appear repeatedly. Such an occurrence is certainly unacceptable. This can happen rapidly with some methods.
4. The random numbers should be replicable. Given the starting point (or conditions) it should be possible to generate the same set of random numbers, completely independent of the system that is being simulated. This is helpful for debugging purposes and is a means of facilitating comparisons between systems (see Chapter 12). For the same reasons, it should be possible to easily specify different starting points, widely separated, within the sequence.
5. Most important, the generated random numbers should closely approximate the ideal statistical properties of uniformity and independence.

Inventing techniques that seem to generate random numbers is easy; inventing techniques that really do produce sequences that appear to be independent, uniformly distributed random numbers is incredibly difficult. There is now a vast literature and rich theory on the topic, and many hours of testing have been devoted to establishing the properties of various generators. Even when a technique is known to be theoretically sound, it is seldom easy to implement it in a way that will be fast and portable. The goal of this chapter is to make the reader aware of the central issues in random-number generation, to enhance understanding and to show some of the techniques that are used by those working in this area.

7.3 TECHNIQUES FOR GENERATING RANDOM NUMBERS

The linear congruential method of Section 7.3.1 is the most widely used technique for generating random numbers, so we describe it in detail. We also report an extension of this method that yields sequences with a longer period. Many other methods have been proposed, and they are reviewed in Bratley, Fox, and Schrage [1996], Law and Kelton [2000], and Ripley [1987].

7.3.1 Linear Congruential Method

The linear congruential method, initially proposed by Lehmer [1951], produces a sequence of integers, $X_1, X_2, \ldots$ between zero and $m - 1$ by following a recursive relationship:

$$X_{i+1} = (aX_i + c) \bmod m, \quad i = 0, 1, 2, \ldots \tag{7.1}$$

The initial value X_0 is called the seed, a is called the multiplier, c is the increment, and m is the modulus. If $c \neq 0$ in Equation (7.1), then the form is called the *mixed congruential method*. When $c = 0$, the form is known as the *multiplicative congruential method*. The selection of the values for a, c, m, and X_0 drastically

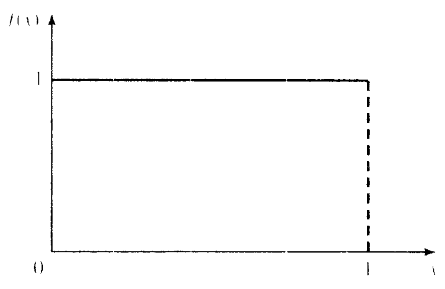

Figure 7.1 pdf for random numbers.

Some consequences of the uniformity and independence properties are the following:

1. If the interval [0, 1] is divided into n classes, or subintervals of equal length, the expected number of observations in each interval is N/n, where N is the total number of observations.
2. The probability of observing a value in a particular interval is independent of the previous values drawn.

7.2 GENERATION OF PSEUDO-RANDOM NUMBERS

Notice that the title of this section has the word "pseudo" in it. "Pseudo" means false, so false random numbers are being generated! In this instance, "pseudo" is used to imply that the very act of generating random numbers by a known method removes the potential for true randomness. If the method is known, the set of random numbers can be replicated. Then an argument can be made that the numbers are not truly random. The goal of any generation scheme, however, is to produce a sequence of numbers between 0 and 1 that simulates, or imitates, the ideal properties of uniform distribution and independence as closely as possible.

To be sure, in the generation of pseudo-random numbers, certain problems or errors can occur. These errors, or departures from ideal randomness, are all related to the properties stated previously. Some examples of such departures include the following:

1. The generated numbers might not be uniformly distributed.
2. The generated numbers might be discrete-valued instead of continuous-valued.
3. The mean of the generated numbers might be too high or too low.
4. The variance of the generated numbers might be too high or too low.
5. There might be dependence. The following are examples:
   - **(a)** autocorrelation between numbers;
   - **(b)** numbers successively higher or lower than adjacent numbers;
   - **(c)** several numbers above the mean followed by several numbers below the mean.

Departures from uniformity and independence for a particular generation scheme often can be detected by such tests as those described in Section 7.4. If such departures are detected, the generation scheme should be dropped in favor of an acceptable generator. Generators that pass the tests in Section 7.4 and tests even more stringent have been developed; thus, there is no excuse for using a generator that has been found to be defective.

affects the statistical properties and the cycle length. Variations of Equation (7.1) are quite common in the computer generation of random numbers. An example will illustrate how this technique operates.

Example 7.1

Use the linear congruential method to generate a sequence of random numbers with $X_0 = 27$, $a = 17$, $c = 43$, and $m = 100$. Here, the integer values generated will all be between zero and 99 because of the value of the modulus. Also, notice that random integers are being generated rather than random numbers. These random integers should appear to be uniformly distributed on the integers zero to 99. Random numbers between zero and 1 can be generated by

$$R_i = \frac{X_i}{m}, \qquad i = 1, 2, \ldots \tag{7.2}$$

The sequence of X_i and subsequent R_i values is computed as follows:

$$\begin{aligned}
X_0 &= 27 \\
X_1 &= (17 \cdot 27 + 43) \bmod 100 = 502 \bmod 100 = 2 \\
R_1 &= \frac{2}{100} = 0.02 \\
X_2 &= (17 \cdot 2 + 43) \bmod 100 = 77 \bmod 100 = 77 \\
R_2 &= \frac{77}{100} = 0.77 \\
X_3 &= (17 \cdot 77 + 43) \bmod 100 = 1352 \bmod 100 = 52 \\
R_3 &= \frac{52}{100} = 0.52 \\
&\vdots
\end{aligned}$$

Recall that $a = b \bmod m$ provided that $(b - a)$ is divisible by m with no remainder. Thus, $X_1 = 502 \bmod 100$, but 502/100 equals 5 with a remainder of 2, so that $X_1 = 2$. In other words, $(502 - 2)$ is evenly divisible by $m = 100$, so $X_1 = 502$ "reduces" to $X_1 = 2 \bmod 100$. (A shortcut for the modulo, or reduction operation for the case $m = 10^b$, a power of 10, is illustrated in Example 7.3.)

The ultimate test of the linear congruential method, as of any generation scheme, is how closely the generated numbers $R_1, R_2, \ldots$ approximate uniformity and independence. There are, however, several secondary properties that must be considered. These include *maximum density* and *maximum period*.

First, notice that the numbers generated from Equation (7.2) assume values only from the set $I = \{0, 1/m, 2/m, \ldots, (m-1)/m\}$, because each X_i is an integer in the set $\{0, 1, 2, \ldots, m-1\}$. Thus, each R_i is discrete on I, instead of continuous on the interval $[0, 1]$. This approximation appears to be of little consequence if the modulus m is a very large integer. (Values such as $m = 2^{31} - 1$ and $m = 2^{48}$ are in common use in generators appearing in many simulation languages.) By maximum density is meant that the values assumed by R_i, $i = 1, 2, \ldots$, leave no large gaps on $[0, 1]$.

Second, to help achieve maximum density, and to avoid cycling (i.e., recurrence of the same sequence of generated numbers) in practical applications, the generator should have the largest possible period. Maximal period can be achieved by the proper choice of a, c, m, and X_0 [Fishman, 1978; Law and Kelton, 2000].

- For m a power of 2, say $m = 2^b$, and $c \neq 0$, the longest possible period is $P = m = 2^b$, which is achieved whenever c is relatively prime to m (that is, the greatest common factor of c and m is 1) and $a = 1 + 4k$, where k is an integer.

- For m a power of 2, say $m = 2^b$, and $c = 0$, the longest possible period is $P = m/4 = 2^{b-2}$, which is achieved if the seed X_0 is odd and if the multiplier, a, is given by $a = 3 + 8k$ or $a = 5 + 8k$, for some $k = 0, 1, \ldots$.
- For m a prime number and $c = 0$, the longest possible period is $P = m - 1$, which is achieved whenever the multiplier, a, has the property that the smallest integer k such that $a^k - 1$ is divisible by m is $k = m - 1$.

Example 7.2

Using the multiplicative congruential method, find the period of the generator for $a = 13$, $m = 2^6 = 64$ and $X_0 = 1, 2, 3$, and 4. The solution is given in Table 7.1. When the seed is 1 or 3, the sequence has period 16. However, a period of length eight is achieved when the seed is 2 and a period of length four occurs when the seed is 4.

In Example 7.2, $m = 2^6 = 64$ and $c = 0$. The maximal period is therefore $P = m/4 = 16$. Notice that this period is achieved by using odd seeds, $X_0 = 1$ and $X_0 = 3$: even seeds, $X_0 = 2$ and $X_0 = 4$, yield the periods eight and four, respectively, both less than the maximum. Notice that $a = 13$ is of the form $5 + 8k$ with $k = 1$, as is required to achieve maximal period.

When $X_0 = 1$, the generated sequence assumes values from the set $\{1, 5, 9, 13, \ldots, 53, 57, 61\}$. The "gaps" in the sequence of generated random numbers, R_i, are quite large (i.e., the gap is $5/64 - 1/64$ or 0.0625). Such a gap gives rise to concern about the density of the generated sequence.

The generator in Example 7.2 is not viable for any application—its period is too short, and its density is insufficient. However, the example shows the importance of properly choosing a, c, m, and X_0.

Speed and efficiency in using the generator on a digital computer is also a selection consideration. Speed and efficiency are aided by use of a modulus, m, which is either a power of 2 or close to a power of 2. Since most digital computers use a binary representation of numbers, the modulo, or remaindering, operation of Equation (7.1) can be conducted efficiently when the modulo is a power of 2 (i.e., $m = 2^b$). After ordinary arithmetic yields a value for $aX_i + c$, X_{i+1} is obtained by dropping the leftmost binary digits in $aX_i + c$ and

Table 7.1 Period Determination Using Various Seeds

i	X_i	X_i	X_i	X_i
0	1	2	3	4
1	13	26	39	52
2	41	18	59	36
3	21	42	63	20
4	17	34	51	4
5	29	58	23	
6	57	50	43	
7	37	10	47	
8	33	2	35	
9	45		7	
10	9		27	
11	53		31	
12	49		19	
13	61		55	
14	25		11	
15	5		15	
16	1		3	

then using only the b rightmost binary digits. The following example illustrates, by analogy, this operation using $m = 10^b$, because most human beings think in decimal representation.

Example 7.3

Let $m = 10^2 = 100$, $a = 19$, $c = 0$, and $X_0 = 63$, and generate a sequence of random integers using Equation (7.1).

$$\begin{aligned}
X_0 &= 63\\
X_1 &= (19)(63) \bmod 100 = 1197 \bmod 100 = 97\\
X_2 &= (19)(97) \bmod 100 = 1843 \bmod 100 = 43\\
X_3 &= (19)(43) \bmod 100 = 817 \bmod 100 = 17\\
&\vdots
\end{aligned}$$

When m is a power of 10, say $m = 10^b$, the modulo operation is accomplished by saving the b rightmost (decimal) digits. By analogy, the modulo operation is most efficient for binary computers when $m = 2^b$ for some $b > 0$.

Example 7.4

The last example in this section is in actual use. It has been extensively tested [Learmonth and Lewis, 1973; Lewis *et al.*, 1969]. The values for a, c, and m have been selected to ensure that the characteristics desired in a generator are most likely to be achieved. By changing X_0, the user can control the repeatability of the stream.

Let $a = 7^5 = 16{,}807$; $m = 2^{31} - 1 = 2{,}147{,}483{,}647$ (a prime number); and $c = 0$. These choices satisfy the conditions that insure a period of $P = m - 1$ (well over 2 billion). Further, specify the seed $X_0 = 123{,}457$. The first few numbers generated are as follows:

$$\begin{aligned}
X_1 &= 7^5(123{,}457) \bmod (2^{31} - 1) = 2{,}074{,}941{,}799 \bmod (2^{31} - 1)\\
X_1 &= 2{,}074{,}941{,}799\\
R_1 &= \frac{X_1}{2^{31}} = 0.9662\\
X_2 &= 7^5(2{,}074{,}941{,}799) \bmod (2^{31} - 1) = 559{,}872{,}160\\
R_2 &= \frac{X_2}{2^{31}} = 0.2607\\
X_3 &= 7^5(559{,}872{,}160) \bmod (2^{31} - 1) = 1{,}645{,}535{,}613\\
R_3 &= \frac{X_3}{2^{31}} = 0.7662\\
&\vdots
\end{aligned}$$

Notice that this routine divides by $m + 1$ instead of m; however, for such a large value of m, the effect is negligible.

7.3.2 Combined Linear Congruential Generators

As computing power has increased, the complexity of the systems that we are able to simulate has also increased. A random-number generator with period $2^{31} - 1 \approx 2 \times 10^9$, such as the popular generator described in Example 7.4, is no longer adequate for all applications. Examples include the simulation of highly reliable systems, in which hundreds of thousands of elementary events must be simulated to observe even a single failure event, and the simulation of complex computer networks, in which thousands of users are executing hundreds of programs. An area of current research is the deriving of generators with substantially longer periods.

One fruitful approach is to combine two or more multiplicative congruential generators in such a way that the combined generator has good statistical properties and a longer period. The following result from L'Ecuyer [1988] suggests how this can be done:

If $W_{i,1}, W_{i,2}, \ldots, W_{i,k}$ are any independent, discrete-valued random variables (not necessarily identically distributed), but one of them, say $W_{i,1}$, is uniformly distributed on the integers from 0 to $m_1 - 2$, then

$$W_i = \left(\sum_{j=1}^{k} W_{i,j} \right) \bmod m_1 - 1$$

is uniformly distributed on the integers from 0 to $m_1 - 2$.

To see how this result can be used to form combined generators, let $X_{i,1}, X_{i,2}, \ldots, X_{i,k}$ be the ith output from k different multiplicative congruential generators, where the jth generator has prime modulus m_j and the multiplier a_j is chosen so that the period is $m_j - 1$. Then the jth generator is producing integers $X_{i,j}$ that are approximately uniformly distributed on the integers from 1 to $m_j - 1$, and $W_{i,j} = X_{i,j} - 1$ is approximately uniformly distributed on the integers from 0 to $m_j - 2$. L'Ecuyer [1988] therefore suggests combined generators of the form

$$X_i = \left(\sum_{j=1}^{k} (-1)^{j-1} X_{i,j} \right) \bmod m_1 - 1$$

with

$$R_i = \begin{cases} \dfrac{X_i}{m_1}, & X_i > 0 \\[2ex] \dfrac{m_1 - 1}{m_1}, & X_i = 0 \end{cases}$$

Notice that the "$(-1)^{j-1}$" coefficient implicitly performs the subtraction $X_{i,1} - 1$; for example, if $k = 2$ then $(-1)^0(X_{i,1} - 1) - (-1)^1(X_{i,2} - 1) = \sum_{j=1}^{2} (-1)^{j-1} X_{i,j}$.

The maximum possible period for such a generator is

$$P = \frac{(m_1 - 1)(m_2 - 1) \cdots (m_k - 1)}{2^{k-1}}$$

which is achieved by the generator described in the next example.

Example 7.5

For 32-bit computers, L'Ecuyer [1988] suggests combining $k = 2$ generators with $m_1 = 2{,}147{,}483{,}563$, $a_1 = 40{,}014$, $m_2 = 2{,}147{,}483{,}399$ and $a_2 = 40{,}692$. This leads to the following algorithm:

1. Select seed $X_{1,0}$ in the range [1,2,14,74,83,562] for the first generator, and seed $X_{2,0}$ in the range [1,2,14,74,83,398] for the second.
   Set $j = 0$.
2. Evaluate each individual generator.

$$X_{1,j+1} = 40{,}014\, X_{1,j} \bmod 2{,}147{,}483{,}563$$
$$X_{2,j+1} = 40{,}692\, X_{2,j} \bmod 2{,}147{,}483{,}399$$

3. Set

$$X_{j+1} = (X_{1,j+1} - X_{2,j+1}) \bmod 2{,}147{,}483{,}562$$

4. Return

$$R_{j+1} = \begin{cases} \dfrac{X_{j+1}}{2{,}147{,}483{,}563}, & X_{j+1} > 0 \\[2ex] \dfrac{2{,}147{,}483{,}562}{2{,}147{,}483{,}563}, & X_{j+1} = 0 \end{cases}$$

5. Set $j = j + 1$ and go to step 2.

This combined generator has period $(m_1 - 1)(m_2 - 1)/2 \approx 2 \times 10^{18}$. Perhaps surprisingly, even such a long period might not be adequate for all applications. See L'Ecuyer [1996, 1999] and L'Ecuyer *et al.* [2002] for combined generators with periods as long as $2^{191} \approx 3 \times 10^{57}$.

7.3.3 Random-Number Streams

The seed for a linear congruential random-number generator (seeds, in the case of a combined linear congruential generator) is the integer value X_0 that initializes the random-number sequence. Since the sequence of integers $X_0, X_1, \ldots, X_P, X_0, X_1, \ldots$ produced by a generator repeats, any value in the sequence could be used to "seed" the generator.

For a linear congruential generator, a random-number *stream* is nothing more than a convenient way to refer to a starting seed taken from the sequence $X_0, X_1, \ldots, X_P$ (for a combined generator, starting seeds for all of the basic generators are required); typically these starting seeds are far apart in the sequence. For instance, if the streams are b values apart, then stream i could be defined by starting seed

$$S_i = X_{b(i-1)}$$

for $i = 1, 2, \ldots, \lfloor P/b \rfloor$. Values of $b = 100{,}000$ were common in older generators, but values as large as $b = 10^{37}$ are in use in modern combined linear congruential generators. (See, for instance, L'Ecuyer *et al.* [2002] for the implementation of such a generator.) Thus, a single random-number generator with k streams acts like k distinct virtual random-number generators, provided that the current value of seed for each stream is maintained. Exercise 21 illustrates one way to create streams that are widely separated in the random-number sequence.

In Chapter 12, we will consider the problem of comparing two or more alternative systems via simulation, and we will show that there are advantages to dedicating portions of the pseudorandom number sequence to the same purpose in each of the simulated systems. For instance, in comparing the efficiency of several queueing systems, a fairer comparison will be achieved if all of the simulated systems experience exactly the same sequence of customer arrivals. Such synchronization can be achieved by assigning a specific stream to generate arrivals in each of the queueing simulations. If the starting seeds for the streams are spaced far enough apart, then this has the same effect as having a distinct random-number generator whose only purpose is to generate customer arrivals.

7.4 TESTS FOR RANDOM NUMBERS

The desirable properties of random numbers—uniformity and independence—were discussed in Section 7.1. To check on whether these desirable properties have been achieved, a number of tests can be performed.

(Fortunately, the appropriate tests have already been conducted for most commercial simulation software.) The tests can be placed in two categories, according to the properties of interest: uniformity, and independence. A brief description of two types of tests is given in this chapter:

1. *Frequency test.* Uses the Kolmogorov–Smirnov or the chi-square test to compare the distribution of the set of numbers generated to a uniform distribution.
2. *Autocorrelation test.* Tests the correlation between numbers and compares the sample correlation to the expected correlation, zero.

In testing for uniformity, the hypotheses are as follows:

$$H_0: \quad R_i \sim U[0, 1]$$
$$H_1: \quad R_i \not\sim U[0, 1]$$

The null hypothesis, H_0, reads that the numbers are distributed uniformly on the interval [0, 1]. Failure to reject the null hypothesis means that evidence of nonuniformity has not been detected by this test. This does not imply that further testing of the generator for uniformity is unnecessary.

In testing for independence, the hypotheses are as follows:

$$H_0: \quad R_i \sim \text{independently}$$
$$H_1: \quad R_i \not\sim \text{independently}$$

This null hypothesis, H_0, reads that the numbers are independent. Failure to reject the null hypothesis means that evidence of dependence has not been detected by this test. This does not imply that further testing of the generator for independence is unnecessary.

For each test, a level of significance α must be stated. The level α is the probability of rejecting the null hypothesis when the null hypothesis is true:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

The decision maker sets the value of a for any test. Frequently, α is set to 0.01 or 0.05.

If several tests are conducted on the same set of numbers, the probability of rejecting the null hypothesis on at least one test, by chance alone [i.e., making a Type I (α) error], increases. Say that $\alpha = 0.05$ and that five different tests are conducted on a sequence of numbers. The probability of rejecting the null hypothesis on at least one test, by chance alone, could be as large as 0.25.

Similarly, if one test is conducted on many sets of numbers from a generator, the probability of rejecting the null hypothesis on at least one test by chance alone [i.e., making a Type I (α) error], increases as more sets of numbers are tested. For instance, if 100 sets of numbers were subjected to the test, with $\alpha = 0.05$, it would be expected that five of those tests would be rejected by chance alone. If the number of rejections in 100 tests is close to 100α, then there is no compelling reason to discard the generator. The concept discussed in this and the preceding paragraph is discussed further at the conclusion of Example 7.8.

If one of the well-known simulation languages or random-number generators is used, it is probably unnecessary to apply the tests just mentioned and described in Sections 7.4.1 and 7.4.2. However, random-number generators frequently are added to software that is not specifically developed for simulation, such as spreadsheet programs, symbolic/numerical calculators, and programming languages. If the generator that is at hand is not explicitly known or documented, then the tests in this chapter should be applied to many samples of numbers from the generator. Some additional tests that are commonly used, but are not covered here, are Good's serial test for sampling numbers [1953, 1967], the median-spectrum test [Cox and Lewis, 1966; Durbin, 1967], the runs test [Law and Kelton 2000] and a variance heterogeneity test [Cox and Lewis, 1966]. Even if a set of numbers passes all the tests, there is no guarantee of randomness; it is always possible that some underlying pattern has gone undetected.

In this book, we emphasize empirical tests that are applied to actual sequences of numbers produced by a generator. Because of the extremely long period of modern pseudo-random-number generators, as described in Section 7.3.2, it is no longer possible to apply these tests to a significant portion of the period of such generators. The tests can be used as a check if one encounters a generator with completely unknown properties (perhaps one that is undocumented and buried deep in a software package), but they cannot be used to establish the quality of a generator throughout its period. Fortunately, there are also families of theoretical tests that evaluate the choices for m, a, and c without actually generating any numbers, the most common being the spectral test. Many of these tests assess how k-tuples of random numbers fill up a k-dimensional unit cube. These tests are beyond the scope of this book; see, for instance, Ripley [1987].

In the examples of tests that follow, the hypotheses are not restated. The hypotheses are as indicated in the foregoing paragraphs. Although few simulation analysts will need to perform these tests, every simulation user should be aware of the qualities of a good random-number generator.

7.4.1 Frequency Tests

A basic test that should always be performed to validate a new generator is the test of uniformity. Two different methods of testing are available. They are the Kolmogorov-Smirnov and the chi-square test. Both of these tests measure the degree of agreement between the distribution of a sample of generated random numbers and the theoretical uniform distribution. Both tests are based on the null hypothesis of no significant difference between the sample distribution and the theoretical distribution.

1. *The Kolmogorov-Smirnov test.* This test compares the continuous cdf, $F(x)$, of the uniform distribution with the empirical cdf, $S_N(x)$, of the sample of N observations. By definition,

$$F(x) = x, \quad 0 \le x \le 1$$

If the sample from the random-number generator is $R_1, R_2, \ldots, R_N$, then the empirical cdf, $S_N(x)$, is defined by

$$S_N(x) = \frac{\text{number of } R_1, R_2, \ldots, R_N \text{ which are} \le x}{N}$$

As N becomes larger, $S_N(x)$ should become a better approximation to $F(x)$, provided that the null hypothesis is true.

In Section 5.6, empirical distributions were described. The cdf of an empirical distribution is a step function with jumps at each observed value. This behavior was illustrated by Example 5.35.

The Kolmogorov-Smirnov test is based on the largest absolute deviation between $F(x)$ and $S_N(x)$ over the range of the random variable—that is, it is based on the statistic

$$D = \max|F(x) - S_N(x)| \tag{7.3}$$

The sampling distribution of D is known; it is tabulated as a function of N in Table A.8. For testing against a uniform cdf, the test procedure follows these steps:

Step 1. Rank the data from smallest to largest. Let $R_{(i)}$ denote the ith smallest observation, so that

$$R_{(1)} \le R_{(2)} \le \cdots \le R_{(N)}$$

Step 2. Compute

$$D^{+} = \max_{1 \le i \le N} \left\{ \frac{i}{N} - R_{(i)} \right\}$$

$$D^{-} = \max_{1 \le i \le N} \left\{ R_{(i)} - \frac{i-1}{N} \right\}$$

Step 3. Compute $D = \max(D^{+}, D^{-})$.

Step 4. Locate in Table A.8 the critical value, D_{α}, for the specified significance level α and the given sample size N.

Step 5. If the sample statistic D is greater than the critical value, D_{α}, the null hypothesis that the data are a sample from a uniform distribution is rejected. If $D \le D_{\alpha}$, conclude that no difference has been detected between the true distribution of $\{R_1, R_2, \ldots, R_N\}$ and the uniform distribution.

Example 7.6

Suppose that the five numbers 0.44, 0.81, 0.14, 0.05, 0.93 were generated, and it is desired to perform a test for uniformity by using the Kolmogorov–Smirnov test with the level of significance $\alpha = 0.05$. First, the numbers must be ranked from smallest to largest. The calculations can be facilitated by use of Table 7.2. The top row lists the numbers from smallest ($R_{(1)}$) to largest ($R_{(5)}$). The computations for D^{+}, namely $i/N - R_{(i)}$, and for D^{-}, namely $R_{(i)} - (i-1)/N$, are easily accomplished by using Table 7.2. The statistics are computed as $D^{+} = 0.26$ and $D^{-} = 0.21$. Therefore, $D = \max\{0.26, 0.21\} = 0.26$. The critical value of D, obtained from Table A.8 for $\alpha = 0.05$ and $N = 5$, is 0.565. Since the computed value, 0.26, is less than the tabulated critical value, 0.565, the hypothesis that the distribution of the generated numbers is the uniform distribution is not rejected.

The calculations in Table 7.2 are illustrated in Figure 7.2, where the empirical cdf, $S_N(x)$, is compared to the uniform cdf, $F(x)$. It can be seen that D^{+} is the largest deviation of $S_N(x)$ above $F(x)$, and that D^{-} is the largest deviation of $S_N(x)$ below $F(x)$. For example, at $R_{(3)}$, the value of D^{+} is given by $3/5 - R_{(3)} = 0.60 - 0.44 = 0.16$, and that of D^{-} is given by $R_{(3)} - 2/5 = 0.44 - 0.40 = 0.04$. Although the test statistic D is defined by Equation (7.3) as the maximum deviation over all x, it can be seen from Figure 7.2 that the maximum deviation will always occur at one of the jump points $R_{(1)}, R_{(2)}, \ldots$; thus, the deviation at other values of x need not be considered.

2. *The chi-square test.* The chi-square test uses the sample statistic

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed number in the ith class, E_i is the expected number in the ith class, and n is the number of classes. For the uniform distribution, E_i, the expected number in each class is given by

$$E_i = \frac{N}{n}$$

Table 7.2 Calculations for Kclmogorov–Smirnov Test

$R_{(i)}$	0.05	0.14	0.44	0.81	0.93
i/N	0.20	0.40	0.60	0.80	1.00
$i/N - R_{(i)}$	0.15	0.26	0.16	—	0.07
$R_{(i)} - (i-1)/N$	0.05	—	0.04	0.21	0.13

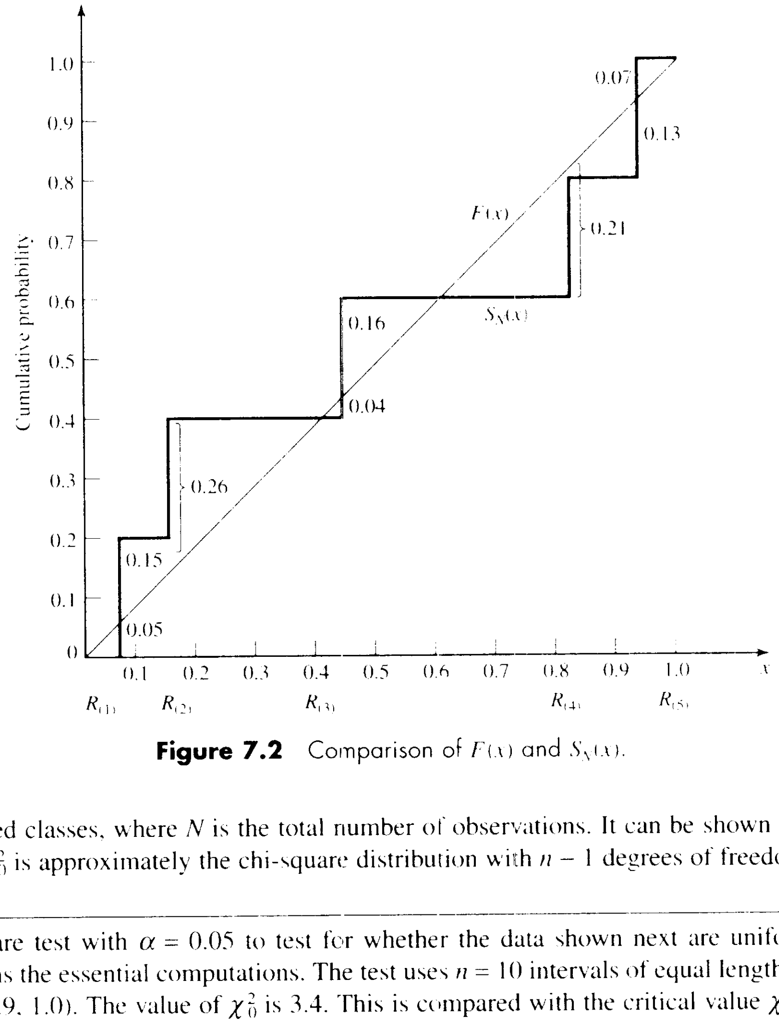

Figure 7.2 Comparison of $F(x)$ and $S_N(x)$.

for equally spaced classes, where N is the total number of observations. It can be shown that the sampling distribution of χ_0^2 is approximately the chi-square distribution with $n - 1$ degrees of freedom.

Example 7.7

Use the chi-square test with $\alpha = 0.05$ to test for whether the data shown next are uniformly distributed. Table 7.3 contains the essential computations. The test uses $n = 10$ intervals of equal length, namely $[0, 0.1)$, $[0.1, 0.2)$, ..., $[0.9, 1.0)$. The value of χ_0^2 is 3.4. This is compared with the critical value $\chi_{0.05,9}^2 = 16.9$ from Table A.6. Since χ_0^2 is much smaller than the tabulated value of $\chi_{0.05,9}^2$, the null hypothesis of a uniform distribution is not rejected.

0.34	0.90	0.25	0.89	0.87	0.44	0.12	0.21	0.46	0.67
0.83	0.76	0.79	0.64	0.70	0.81	0.94	0.74	0.22	0.74
0.96	0.99	0.77	0.67	0.56	0.41	0.52	0.73	0.99	0.02
0.47	0.30	0.17	0.82	0.56	0.05	0.45	0.31	0.78	0.05
0.79	0.71	0.23	0.19	0.82	0.93	0.65	0.37	0.39	0.42
0.99	0.17	0.99	0.46	0.05	0.66	0.10	0.42	0.18	0.49
0.37	0.51	0.54	0.01	0.81	0.28	0.69	0.34	0.75	0.49
0.72	0.43	0.56	0.97	0.30	0.94	0.96	0.58	0.73	0.05
0.06	0.39	0.84	0.24	0.40	0.64	0.40	0.19	0.79	0.62
0.18	0.26	0.97	0.88	0.64	0.47	0.60	0.11	0.29	0.78

Different authors have offered considerations concerning the application of the χ^2 test. In the application to a data set the size of that in Example 7.7, the considerations do not apply—that is, if 100 values are in the

Table 7.3 Computations for Chi-Square Test

Interval	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	8	10	−2	4	0.4
2	8	10	−2	4	0.4
3	10	10	0	0	0.0
4	9	10	−1	1	0.1
5	12	10	2	4	0.4
6	8	10	−2	4	0.4
7	10	10	0	0	0.0
8	14	10	4	16	1.6
9	10	10	0	0	0.0
10	11	10	1	1	0.1
	100	100	0		3.4

sample and from 5 to 10 intervals of equal length are used, the test will be acceptable. In general, it is recommended that n and N be chosen so that each $E_i \geq 5$.

Both the Kolmogorov–Smirnov test and the chi-square test are acceptable for testing the uniformity of a sample of data, provided that the sample size is large. However, the Kolmogorov–Smirnov test is the more powerful of the two and is recommended. Furthermore, the Kolmogorov–Smirnov test can be applied to small sample sizes, whereas the chi-square is valid only for large samples, say $N \geq 50$.

Imagine a set of 100 numbers which are being tested for independence, one where the first 10 values are in the range 0.01–0.10, the second 10 values are in the range 0.11–0.20, and so on. This set of numbers would pass the frequency tests with ease, but the ordering of the numbers produced by the generator would not be random. The test in the next section of this chapter is concerned with the independence of random numbers that are generated.

7.4.2 Tests for Autocorrelation

The tests for autocorrelation are concerned with the dependence between numbers in a sequence. As an example, consider the following sequence of numbers:

0.12	0.01	0.23	0.28	0.89	0.31	0.64	0.28	0.83	0.93
0.99	0.15	0.33	0.35	0.91	0.41	0.60	0.27	0.75	0.88
0.68	0.49	0.05	0.43	0.95	0.58	0.19	0.36	0.69	0.87

From a visual inspection, these numbers appear random, and they would probably pass all the tests presented to this point. However, an examination of the 5th, 10th, 15th (every five numbers beginning with the fifth), and so on, indicates a very large number in that position. Now, 30 numbers is a rather small sample size on which to reject a random number generator, but the notion is that numbers in the sequence might be related. In this particular section, a method for discovering whether such a relationship exists is described. The relationship would not have to be all high numbers. It is possible to have all low numbers in the locations being examined, or the numbers could alternate from very high to very low.

The test to be described shortly requires the computation of the autocorrelation between every m numbers (m is also known as the lag), starting with the ith number. Thus, the autocorrelation ρ_{im} between the following numbers would be of interest: R_i, R_{i+m}, R_{i+2m}, ..., $R_{i+(M+1)m}$. The value M is the largest integer such

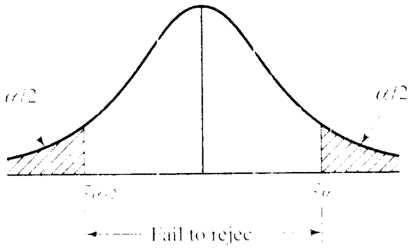

Figure 7.3 Failure to reject hypothesis.

that $i + (M + 1)m \leq N$, where N is the total number of values in the sequence. (Thus, a subsequence of length $M + 2$ is being tested.)

A nonzero autocorrelation implies a lack of independence, so the following two-tailed test is appropriate:

$$H_0: \quad \rho_{im} = 0$$
$$H_1: \quad \rho_{im} \neq 0$$

For large values of M, the distribution of the estimator of ρ_{im}, denoted $\hat{\rho}_{im}$, is approximately normal if the values $R_i, R_{i+m}, R_{i+2m}, \ldots, R_{i+(M+1)m}$ are uncorrelated. Then the test statistic can be formed as follows:

$$Z_0 = \frac{\hat{\rho}_{im}}{\sigma_{\hat{\rho}_{im}}}$$

which is distributed normally with a mean of zero and a variance of 1, under the assumption of independence, for large M.

The formula for $\hat{\rho}_{im}$, in a slightly different form, and the standard deviation of the estimator, $\sigma_{\hat{\rho}_{im}}$, are given by Schmidt and Taylor [1970] as follows:

$$\hat{\rho}_{im} = \frac{1}{M+1}\left[\sum_{k=0}^{M} R_{i+km} R_{i+(k+1)m}\right] - 0.25$$

and

$$\sigma_{\hat{\rho}_{im}} = \frac{\sqrt{13M + 7}}{12(M + 1)}$$

After computing Z_0, do not reject the null hypothesis of independence if $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ where α is the level of significance and $z_{\alpha/2}$ is obtained from Table A.3. Figure 7.3 illustrates this test.

If $\rho_{im} > 0$, the subsequence is said to exhibit positive autocorrelation. In this case, successive values at lag m have a higher probability than expected of being close in value (i.e., high random numbers in the subsequence followed by high, and low followed by low). On the other hand, if $\rho_{im} < 0$, the subsequence is exhibiting negative autocorrelation, which means that low random numbers tend to be followed by high ones, and vice versa. The desired property, independence (which implies zero autocorrelation), means that there is no discernible relationship of the nature discussed here between successive random numbers at lag m.

Example 7.8

Test for whether the 3rd, 8th, 13th, and so on, numbers in the sequence at the beginning of this section are autocorrelated using $\alpha = 0.05$. Here, $i = 3$ (beginning with the third number), $m = 5$ (every five numbers), $N = 30$ (30 numbers in the sequence), and $M = 4$ (largest integer such that $3 + (M + 1)5 \leq 30$). Then,

$$\hat{\rho}_{35} = \frac{1}{4+1}[(0.23)(0.28) + (0.28)(0.33) + (0.33)(0.27) + (0.27)(0.05)$$
$$+ (0.05)(0.36)] - 0.25$$
$$= -0.1945$$

and

$$\sigma_{\hat{\rho}_{35}} = \frac{\sqrt{13(4)+7}}{12(4+1)} = 0.1280$$

Then, the test statistic assumes the value

$$Z_0 = -\frac{0.1945}{0.1280} = -1.516$$

Now, the critical value from Table A.3 is

$$z_{0.025} = 1.96$$

Therefore, the hypothesis of independence cannot be rejected on the basis of this test.

It can be observed that this test is not very sensitive for small values of M, particularly when the numbers being tested are on the low side. Imagine what would happen if each of the entries in the foregoing computation of $\hat{\rho}_{im}$ were equal to zero. Then $\hat{\rho}_{im}$ would be equal to -0.25 and the calculated Z would have the value of -1.95, not quite enough to reject the hypothesis of independence.

There are many sequences that can be formed in a set of data, given a large value of N. For example, beginning with the first number in the sequence, possibilities include (1) the sequence of all numbers, (2) the sequence formed from the first, third, fifth, ..., numbers, (3) the sequence formed from the first, fourth, ..., numbers, and so on. If $\alpha = 0.05$, there is a probability of 0.05 of rejecting a true hypothesis. If 10 independent sequences are examined, the probability of finding no significant autocorrelation, by chance alone, is $(0.95)^{10}$ or 0.60. Thus, 40% of the time significant autocorrelation would be detected when it does not exist. If α is 0.10 and 10 tests are conducted, there is a 65% chance of finding autocorrelation by chance alone. In conclusion, in "fishing" for autocorrelation by performing numerous tests, autocorrelation might eventually be detected, perhaps by chance alone, even when there is no autocorrelation present.

7.5 SUMMARY

This chapter described the generation of random numbers and the subsequent testing of the generated numbers for uniformity and independence. Random numbers are used to generate random variates, the subject of Chapter 8.

Of the many types of random-number generators available, ones based on the linear congruential method are the most widely used, but they are being replaced by combined linear congruential generators. Of the many types of statistical tests that are used in testing random-number generators, two different types are described: one testing for uniformity, and one testing for independence.

The simulation analyst might never work directly with a random-number generator or with the testing of random numbers from a generator. Most computers and simulation languages have routines that generate a random number, or streams of random numbers, for the asking. But even generators that have been used for years, some of which are still in use, have been found to be inadequate. So this chapter calls the simulation analyst's attention to such possibilities, with a warning to investigate and confirm that the generator has been tested thoroughly. Some researchers have attained sophisticated expertise in developing methods for generating and testing random numbers and the subsequent application of these methods. This chapter provides only a basic introduction to the subject matter, more depth and breadth are required for the reader to become a specialist in the area. The bible is Knuth [1998]; see also the reviews in Bratley, Fox, and Schrage [1996], Law and Kelton [2000], L'Ecuyer [1998], and Ripley [1987].

One final caution is due. Even if generated numbers pass all the tests (those covered in this chapter and those mentioned in the chapter), some underlying pattern might have gone undetected without the generator's having been rejected as faulty. However, the generators available in widely used simulation languages have been extensively tested and validated.

REFERENCES

BRATLEY, P., B. L. FOX, AND L. E. SCHRAGE [1996], *A Guide to Simulation*, 2d ed., Springer-Verlag, New York.

COX, D. R., AND P. A. W. LEWIS [1966], *The Statistical Analysis of Series of Events*, Methuen, London.

DURBIN, J. [1967], "Tests of Serial Independence Based on the Cumulated Periodogram," *Bulletin of the International Institute of Statistics*.

FISHMAN, G. S. [1978], *Principles of Discrete Event Simulation*, Wiley, New York.

GOOD, I. J. [1953], "The Serial Test for Sampling Numbers and Other Tests of Randomness," *Proceedings of the Cambridge Philosophical Society*, Vol. 49, pp. 276–284.

GOOD, I. J. [1967], "The Generalized Serial Test and the Binary Expansion of 4," *Journal of the Royal Statistical Society*, Ser. A, Vol. 30, No. 1, pp. 102–107.

KNUTH, D. W. [1998], *The Art of Computer Programming: Vol. 2, Semi-numerical Algorithms*, 2d ed., Addison Wesley, Reading, MA.

LAW, A. M., AND W. D. KELTON [2000], *Simulation Modeling & Analysis*, 3d ed., McGraw-Hill, New York.

LEARMONTH, G. P., AND P. A. W. LEWIS [1973], "Statistical Tests of Some Widely Used and Recently Proposed Uniform Random Number Generators," *Proceedings of the Conference on Computer Science and Statistics: Seventh Annual Symposium on the Interface*, Western Publishing, North Hollywood, CA, pp. 163–171.

L'ECUYER, P. [1988], "Efficient and Portable Combined Random Number Generators," *Communications of the ACM*, Vol. 31, pp. 742–749, 774.

L'ECUYER, P. [1996], "Combined Multiple Recursive Random Number Generators," *Operations Research*, Vol. 44, pp. 816–822.

L'ECUYER, P. [1998], "Random Number Generation," Chapter 4 in *Handbook of Simulation*, Wiley, New York.

L'ECUYER, P. [1999], "Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators," *Operations Research*, Vol. 47, pp. 159–164.

L'ECUYER, P., R. SIMARD, E. J. CHEN, AND W. D. KELTON [2002], "An Object-Oriented Random-Number Package with Many Long Streams and Substreams," *Operations Research*, Vol. 50, pp. 1073–1075.

LEHMER, D. H. [1951], *Proceedings of the Second Symposium on Large-Scale Digital Computing Machinery*, Harvard University Press, Cambridge, MA.

LEWIS, P. A. W., A. S. GOODMAN, AND J. M. MILLER [1969], "A Pseudo-Random Number Generator for the System/360," *IBM Systems Journal*, Vol. 8, pp. 136–145.

RIPLEY, B. D. [1987], *Stochastic Simulation*, Wiley, New York.

SCHMIDT, J. W., And R. E. TAYLOR [1970], *Simulation and Analysis of Industrial Systems*, Irwin, Homewood, IL.

EXERCISES

1. Describe a procedure to physically generate random numbers on the interval [0, 1] with 2-digit accuracy. (*Hint*: Consider drawing something out of a hat.)

2. List applications, other than systems simulation, for pseudo-random numbers—for example, video gambling games.

3. How could random numbers that are uniform on the interval [0, 1] be transformed into random numbers that are uniform on the interval [−11, 17]? Transformations to more general distributions are described in Chapter 8.

4. Generate random numbers using multiplicative congruential method with $X_0 = 5$, $a = 17$, and $m = 64$.

5. Repeat Exercise 4 with $X_0 = 6$, 7, and 8.

6. Generate four-digit random numbers by linear congruential method with $X_0 = 21$, $a = 34$, and $c = 7$.

7. The sequence of numbers 0.54, 0.73, 0.98, 0.11, and 0.68 has been generated. Use the Kolmogorov–Smirnov test with $\alpha = 0.05$ to learn whether the hypothesis that the numbers are uniformly distributed on the interval [0, 1] can be rejected.

8. Generate 1000 random numbers between 0 and 99 using Excel. Conduct chi-square test with $\alpha = 0.05$ and verify whether the numbers are uniformly distributed.

9. Figure out whether these linear congruential generators can achieve a maximum period: also, state restrictions on X_0 to obtain this period

(a) the mixed congruential method with

$$a = 2,814,749,767,109$$
$$c = 59,482,661,568,307$$
$$m = 2^{48}$$

(b) the multiplicative congruential generator with

$$a = 69,069$$
$$c = 0$$
$$m = 2^{32}$$

(c) the mixed congruential generator with

$$a = 4951$$
$$c = 247$$
$$m = 256$$

(d) the multiplicative congruential generator with

$$a = 6507$$
$$c = 0$$
$$m = 1024$$

10. Use the mixed congruential method to generate a sequence of three two-digit random numbers with $X_0 = 37$ $a = 7$, $c = 29$, and $m = 100$.

11. Additive congruential method employs the following expression to generate random numbers:

$$X_{n+1} = (X_1 + X_n) \bmod m$$

where X_1 to X_n are the seeds and X_{n+1} is the new random number. Assuming $n = 5$, $X_1 = 20$, $X_2 = 82$, $X_3 = 42$, $X_4 = 76$, $X_5 = 59$, and $m = 100$, generate 10 new random numbers.

12. Write a computer program to generate random numbers using additive congruential method given in Exercise 11.

13. If $X_0 = 3579$ in Exercise 9(c), generate the first random number in the sequence. Compute the random number to four-place accuracy.

14. Investigate the random-number generator in a spreadsheet program on a computer to which you have access. In many spreadsheets, random numbers are generated by a function called RAND or @RAND.

(a) Check the user's manual to see whether it describes how the random numbers are generated.
(b) Write macros to conduct each of the tests described in this chapter. Generate 100 sets of random numbers, each set containing 100 random numbers. Perform each test on each set of random numbers. Draw conclusions.

15. Consider the multiplicative congruential generator under the following circumstances.

(a) $a = 11, m = 16, X_0 = 7$
(b) $a = 11, m = 16, X_0 = 8$
(c) $a = 7, m = 16, X_0 = 7$
(d) $a = 7, m = 16, X_0 = 8$

Generate enough values in each case to complete a cycle. What inferences can be drawn? Is maximum period achieved?

16. For 16-bit computers, L'Ecuyer [1988] recommends combining three multiplicative generators, with $m_1 = 32{,}363$, $a_1 = 157$, $m_2 = 31{,}727$, $a_2 = 146$, $m_3 = 31{,}657$, and $a_3 = 142$. The period of this generator is approximately 8×10^{12}. Generate 5 random numbers with the combined generator, using the initial seeds $X_{i,0} = 100, 300, 500$, for the individual generators $i = 1, 2, 3$.

17. Search the web and find various other methods of generating random numbers.

18. Use the principles described in this chapter to develop your own linear congruential random-number generator.

19. Use the principles described in this chapter to develop your own combined linear congruential random number generator.

20. The following is the set of single-digit numbers from a random number generator.

6	7	0	6	9	9	0	6	4	6
4	0	8	2	6	6	1	2	6	8
5	6	0	4	7	1	3	5	0	7
1	4	9	8	6	0	9	6	6	7
1	0	4	7	9	2	0	1	4	8
6	9	7	7	5	4	2	3	3	3
6	0	5	8	2	5	8	8	3	1
4	0	8	1	7	0	0	6	2	8
5	6	0	8	0	6	9	7	0	0
3	1	5	4	3	8	3	3	2	4

Using appropriate test, check whether the numbers are uniformly distributed.

21. In some applications, it is useful to be able to quickly skip ahead in a pseudo-random number sequence without actually generating all of the intermediate values. (a) For a linear congruential generator with $c = 0$, show that $X_{i+n} = (a^n X_i) \bmod m$. (b) Next, show that $(a^n X_i) \bmod m = (a^n \bmod m)X_i \bmod m$ (this result is useful because $a^n \bmod m$ can be precomputed, making it easy to skip ahead n random numbers from any point in the sequence). (c) In Example 7.3, use this result to compute X_5, starting with $X_0 = 63$. Check your answer by computing X_5 in the usual way.

8

Random-Variate Generation

This chapter deals with procedures for sampling from a variety of widely-used continuous and discrete distributions. Previous discussions and examples indicated the usefulness of statistical distributions in modeling activities that are generally unpredictable or uncertain. For example, interarrival times and service times at queues and demands for a product are quite often unpredictable in nature, at least to a certain extent. Usually, such variables are modeled as random variables with some specified statistical distribution, and standard statistical procedures exist for estimating the parameters of the hypothesized distribution and for testing the validity of the assumed statistical model. Such procedures are discussed in Chapter 9.

In this chapter, it is assumed that a distribution has been completely specified, and ways are sought to generate samples from this distribution to be used as input to a simulation model. The purpose of the chapter is to explain and illustrate some widely-used techniques for generating random variates, not to give a state-of-the-art survey of the most efficient techniques. In practice, most simulation modelers will use either existing routines available in programming libraries or the routines built into the simulation language being used. However, some programming languages do not have built-in routines for all of the regularly used distributions, and some computer installations do not have random-variate-generation libraries; in such cases the modeler must construct an acceptable routine. Even though the chance of this happening is small, it is nevertheless worthwhile to understand how random-variate generation occurs.

This chapter discusses the inverse-transform technique and, more briefly, the acceptance–rejection technique and special properties. Another technique, the composition method, is discussed by Devroye [1986], Dagpunar [1988], Fishman [1978], and Law and Kelton [2000]. All the techniques in this chapter assume that a source of uniform [0,1] random numbers, $R_1, R_2, \ldots$, is readily available, where each R_i has pdf

$$f_R(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

and cdf

$$F_R(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$$

Throughout this chapter R and $R_1, R_2, \ldots$ represent random numbers uniformly distributed on [0,1] and generated by one of the techniques in Chapter 7 or taken from a random number table, such as Table A.1.

8.1 INVERSE-TRANSFORM TECHNIQUE

The inverse-transform technique can be used to sample from the exponential, the uniform, the Weibull, the triangular distributions and from empirical distributions. Additionally, it is the underlying principle for sampling from a wide variety of discrete distributions. The technique will be explained in detail for the exponential distribution and then applied to other distributions. Computationally, it is the most straightforward, but not always the most efficient, technique.

8.1.1 Exponential Distribution

The exponential distribution, discussed in Section 5.4, has the probability density function (pdf)

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

and the cumulative distribution function (cdf)

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0 \\ 0 & x < 0 \end{cases}$$

The parameter λ can be interpreted as the mean number of occurrences per time unit. For example, if interarrival times $X_1, X_2, X_3, \ldots$ had an exponential distribution with rate λ, then λ could be interpreted as the mean number of arrivals per time unit, or the arrival rate. Notice that, for any i,

$$E(X_i) = \frac{1}{\lambda}$$

and so $1/\lambda$ is the mean interarrival time. The goal here is to develop a procedure for generating values $X_1, X_2, X_3, \ldots$ that have an exponential distribution.

The inverse-transform technique can be utilized, at least in principle, for any distribution, but it is most useful when the cdf, $F(x)$, is of a form so simple that its inverse, F^{-1}, can be computed easily.[1] One step-by-step procedure for the inverse-transform technique, illustrated by the exponential distribution, consists of the following steps:

Step 1. Compute the cdf of the desired random variable X.

For the exponential distribution, the cdf is $F(x) = 1 - e^{-\lambda x}$, $x \ge 0$.

Step 2. Set $F(X) = R$ on the range of X.

For the exponential distribution, it becomes $1 - e^{-\lambda X} = R$ on the range $x \ge 0$.

[1] The notation F^{-1} denotes the solution of the equation $r = F(x)$ in terms of r; it does not denote $1/F$.

X is a random variable (with the exponential distribution in this case), so $1 - e^{-\lambda X}$ is also a random variable, here called R. As will be shown later, R has a uniform distribution over the interval $[0, 1]$.

Step 3. Solve the equation $F(X) = R$ for X in terms of R.

For the exponential distribution, the solution proceeds as follows:

$$1 - e^{-\lambda X} = R$$

$$e^{-\lambda X} = 1 - R$$

$$-\lambda X = \ln(1 - R)$$

$$X = -\frac{1}{\lambda}\ln(1 - R) \tag{8.1}$$

Equation (8.1) is called a random-variate generator for the exponential distribution. In general, Equation (8.1) is written as $X = F^{-1}(R)$. Generating a sequence of values is accomplished through Step 4.

Step 4. Generate (as needed) uniform random numbers R_1, R_2, R_3, ... and compute the desired random variates by

$$X_i = F^{-1}(R_i)$$

For the exponential case, $F^{-1}(R) = (-1/\lambda)\ln(1 - R)$ by Equation (8.1), so

$$X_i = -\frac{1}{\lambda}\ln(1 - R_i) \tag{8.2}$$

for $i = 1, 2, 3, \ldots$. One simplification that is usually employed in Equation (8.2) is to replace $1 - R_i$ by R_i to yield

$$X_i = -\frac{1}{\lambda}\ln R_i \tag{8.3}$$

This alternative is justified by the fact that both R_i and $1 - R_i$ are uniformly distributed on $[0,1]$.

Example 8.1

Table 8.1 gives a sequence of random numbers from Table A.1 and the computed exponential variates, X_i, given by Equation (8.2) with the value $\lambda = 1$. Figure 8.1(a) is a histogram of 200 values, R_1, R_2, ..., R_{200} from the uniform distribution, and Figure 8.1(b) is a histogram of the 200 values, X_1, X_2, ..., X_{200}, computed by Equation (8.2). Compare these empirical histograms with the theoretical density functions in Figure 8.1(c) and (d). As illustrated here, a histogram is an estimate of the underlying density function. (This fact is used in Chapter 9 as a way to identify distributions.)

Table 8.1 Generation of Exponential Variates X_i with Mean 1, given Random Numbers R_i

i	1	2	3	4	5
R_i	0.1306	0.0422	0.6597	0.7965	0.7696
X_i	0.1400	0.0431	1.078	1.592	1.468

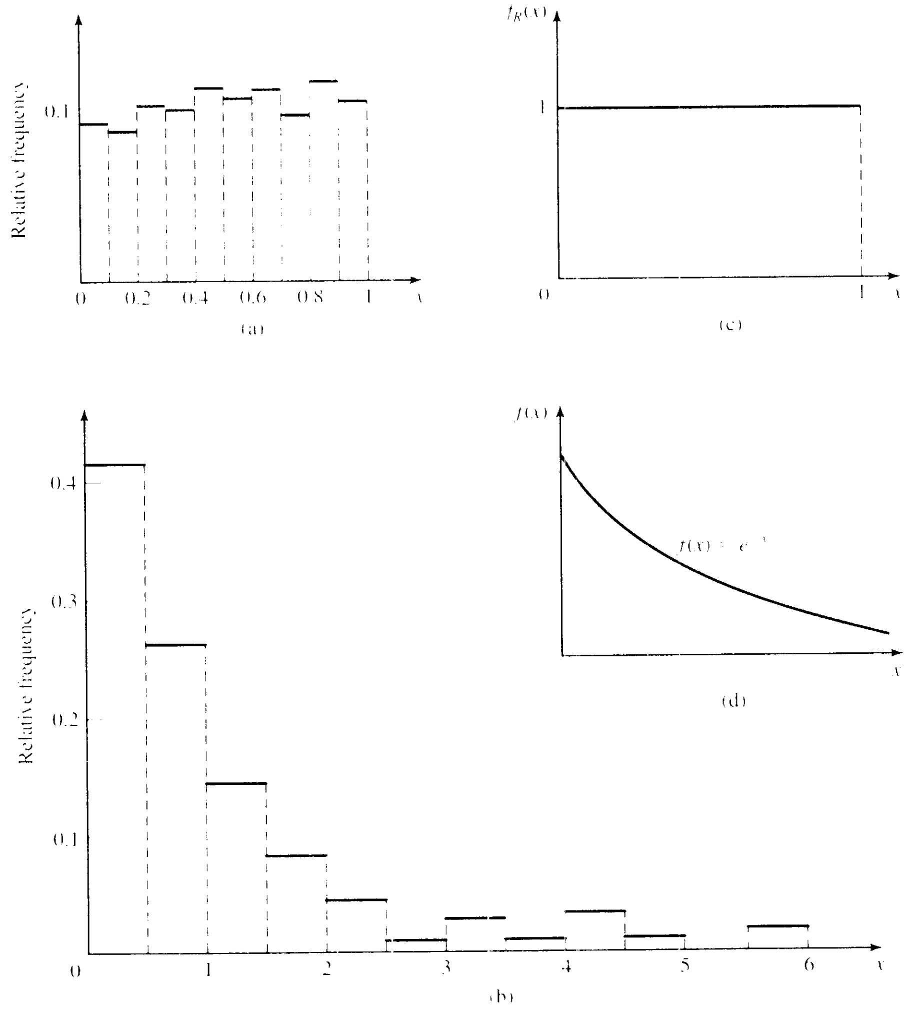

Figure 8.1 (a) Empirical histogram of 200 uniform random numbers; (b) empirical histogram of 200 exponential variates; (c) theoretical uniform density on [0, 1]; (d) theoretical exponential density with mean 1.

Figure 8.2 gives a graphical interpretation of the inverse-transform technique. The cdf shown is $F(x) = 1-e^{-x}$, an exponential distribution with rate $\lambda = 1$. To generate a value X_1 with cdf $F(x)$, a random number R_1 between 0 and 1 is generated, then a horizontal line is drawn from R_1 to the graph of the cdf, then a vertical line is dropped to the x axis to obtain X_1, the desired result. Notice the inverse relation between R_1 and X_1, namely

$$R_1 = 1 - e^{-X_1}$$

and

$$X_1 = -\ln(1 - R_1)$$

In general, the relation is written as

$$R_1 = F(X_1)$$

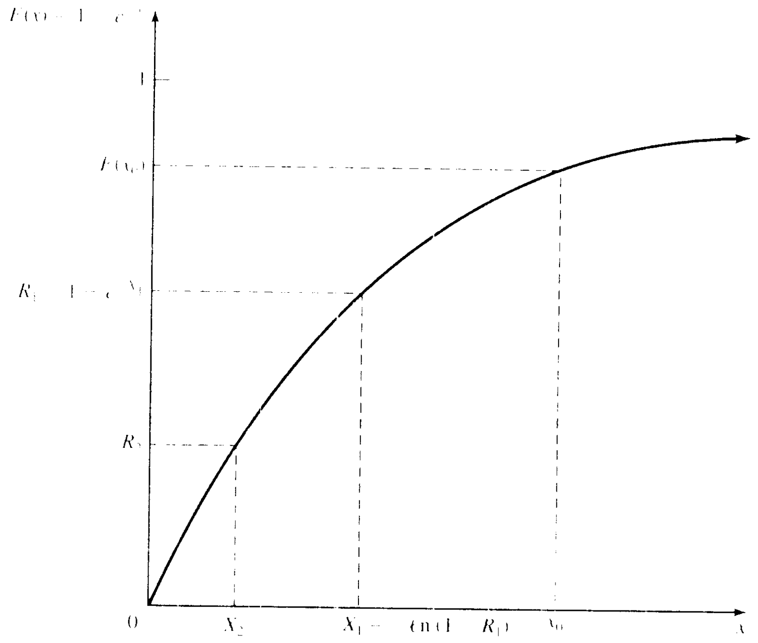

Figure 8.2 Graphical view of the inverse-transform technique.

and

$$X_1 = F^{-1}(R_1)$$

Why does the random variable X_1 generated by this procedure have the desired distribution? Pick a value x_0 and compute the cumulative probability

$$P(X_1 \le x_0) = P(R_1 \le F(x_0)) = F(x_0) \tag{8.4}$$

To see the first equality in Equation (8.4), refer to Figure 8.2, where the fixed numbers x_0 and $F(x_0)$ are drawn on their respective axes. It can be seen that $X_1 \le x_0$ when and only when $R_1 \le F(x_0)$. Since $0 \le F(x_0) \le 1$, the second equality in Equation (8.4) follows immediately from the fact that R_1 is uniformly distributed on $[0, 1]$. Equation (8.4) shows that the cdf of X_1 is F; hence, X_1 has the desired distribution.

8.1.2 Uniform Distribution

Consider a random variable X that is uniformly distributed on the interval $[a, b]$. A reasonable guess for generating X is given by

$$X = a + (b - a)R \tag{8.5}$$

[Recall that R is always a random number on $[0, 1]$.] The pdf of X is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

The derivation of Equation (8.5) follows Steps 1 through 3 of Section 8.1.1:

Step 1. The cdf is given by

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

Step 2. Set $F(X) = (X - a)/(b - a) = R$.

Step 3. Solving for X in terms of R yields $X = a + (b - a)R$, which agrees with Equation (8.5).

8.1.3 Weibull Distribution

The Weibull distribution was introduced in Section 5.4 as a model for *time to failure* for machines or electronic components. When the location parameter ν is set to 0, its pdf is given by Equation (5.47):

$$f(x) = \begin{cases} \dfrac{\beta}{\alpha^{\beta}} x^{\beta-1} e^{-(x/\alpha)^{\beta}}, & x \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\alpha > 0$ and $\beta > 0$ are the scale and shape parameters of the distribution. To generate a Weibull variate, follow Steps 1 through 3 of Section 8.1.1:

Step 1. The cdf is given by $F(X) = 1 - e^{-(x/\alpha)^{\beta}}, x \ge 0$.

Step 2. Set $F(X) = 1 - e^{-(X/\alpha)^{\beta}} = R$.

Step 3. Solving for X in terms of R yields

$$X = \alpha[-\ln(1 - R)]^{1/\beta} \tag{8.6}$$

The derivation of Equation (8.6) is left as Exercise 10 for the reader. By comparing Equations (8.6) and (8.1), it can be seen that, if X is a Weibull variate, then X^{β} is an exponential variate with mean α^{β}. Conversely, if Y is an exponential variate with mean μ, then $Y^{1/\beta}$ is a Weibull variate with shape parameter β and scale parameter $\alpha = \mu^{1/\beta}$.

8.1.4 Triangular Distribution

Consider a random variable X that has pdf

$$f(x) = \begin{cases} x, & 0 \le x \le 1 \\ 2 - x, & 1 < x \le 2 \\ 0, & \text{otherwise} \end{cases}$$

as shown in Figure 8.3. This distribution is called a triangular distribution with endpoints (0, 2) and mode at 1. Its cdf is given by

$$F(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{x^2}{2}, & 0 < x \le 1 \\ 1 - \dfrac{(2-x)^2}{2}, & 1 < x \le 2 \\ 1, & x > 2 \end{cases}$$

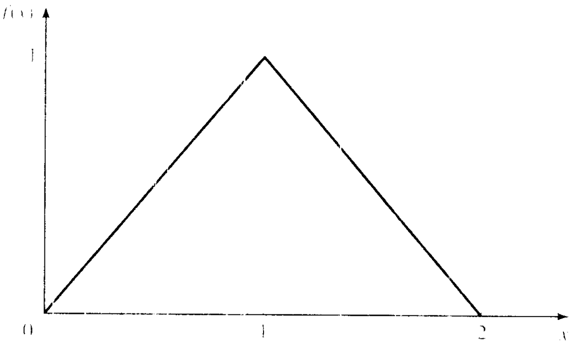

Figure 8.3 Density function for a particular triangular distribution.

For $0 \le X \le 1$,

$$R = \frac{X^2}{2} \tag{8.7}$$

and for $1 \le X \le 2$,

$$R = 1 - \frac{(2-X)^2}{2} \tag{8.8}$$

By Equation (8.7), $0 \le X \le 1$ implies that $0 \le R \le \frac{1}{2}$, in which case $X = \sqrt{2R}$. By Equation (8.8), $1 \le X \le 2$ implies that $\frac{1}{2} \le R \le 1$, in which case $X = 2 - \sqrt{2(1-R)}$. Thus, X is generated by

$$X = \begin{cases} \sqrt{2R}, & 0 \le R \le \frac{1}{2} \\ 2 - \sqrt{2(1-R)}, & \frac{1}{2} < R \le 1 \end{cases} \tag{8.9}$$

Exercises 2, 3, and 4 give the student practice in dealing with other triangular distributions. Notice that, if the pdf and cdf of the random variable X come in parts (i.e., require different formulas over different parts of the range of X), then the application of the inverse-transform technique for generating X will result in separate formulas over different parts of the range of R, as in Equation (8.9). A general form of the triangular distribution was discussed in Section 5.4.

8.1.5 Empirical Continuous Distributions

If the modeler has been unable to find a theoretical distribution that provides a good model for the input data, then it may be necessary to use the empirical distribution of the data. One possibility is to simply resample the observed data itself. This is known as using the *empirical distribution*, and it makes particularly good sense when the input process is known to take on a finite number of values. See Section 8.1.7 for an example of this type of situation and for a method for generating random inputs.

On the other hand, if the data are drawn from what is believed to be a continuous-valued input process, then it makes sense to interpolate between the observed data points to fill in the gaps. This section describes a method for defining and generating data from a continuous empirical distribution.

Example 8.2

Five observations of fire-crew response times (in minutes) to incoming alarms have been collected to be used in a simulation investigating possible alternative staffing and crew-scheduling policies. The data are

2.76 1.83 0.80 1.45 1.24

Before collecting more data, it is desired to develop a preliminary simulation model that uses a response-time distribution based on these five observations. Thus, a method for generating random variates from the response-time distribution is needed. Initially, it will be assumed that response times X have a range $0 \leq X \leq c$, where c is unknown, but will be estimated by $\hat{c} = \max\{X_i : i = 1, \ldots, n\} = 2.76$, where $\{X_i, i = 1, \ldots, n\}$ are the raw data and $n = 5$ is the number of observations.

Arrange the data from smallest to largest and let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ denote these sorted values. The smallest possible value is believed to be 0, so define $x_{(0)} = 0$. Assign the probability $1/n = 1/5$ to each interval

Table 8.2 Summary of Fire-Crew Response-Time Data

i	*Interval* $x_{(i-1)} < x \leq x_{(i)}$	*Probability* $1/n$	*Cumulative Probability,* i/n	*Slope* a_i
1	$0.0 < x \leq 0.80$	0.2	0.2	4.00
2	$0.80 < x \leq 1.24$	0.2	0.4	2.20
3	$1.24 < x \leq 1.45$	0.2	0.6	1.05
4	$1.45 < x \leq 1.83$	0.2	0.8	1.90
5	$1.83 < x \leq 2.76$	0.2	1.0	4.65

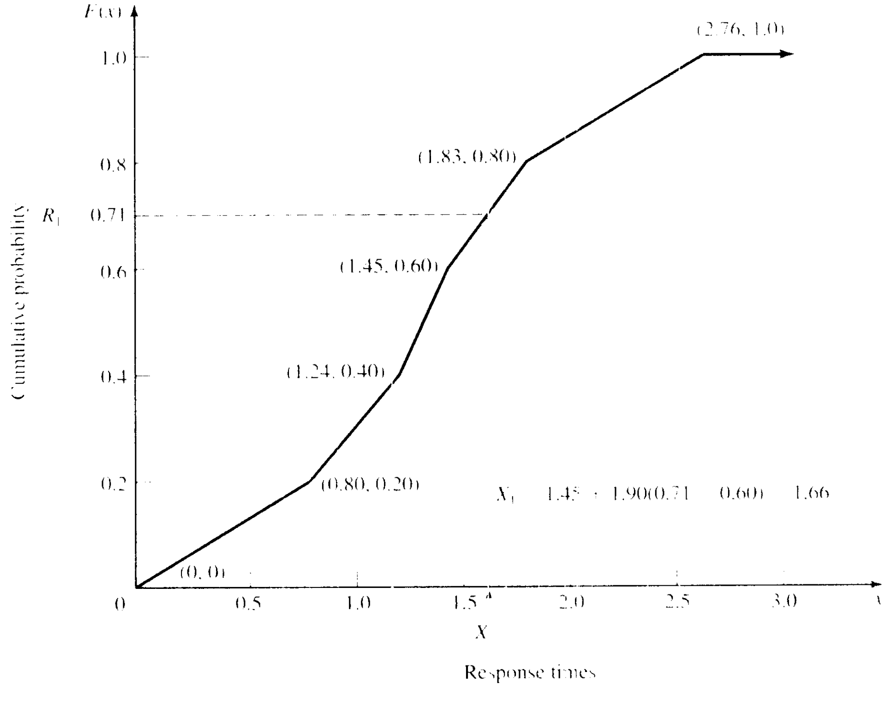

Figure 8.4 Empirical cdf of fire-crew response times.

$x_{(i-1)} < x \leq x_{(i)}$, as shown in Table 8.2. The resulting empirical cdf, $\hat{F}(x)$, is illustrated in Figure 8.4. The slope of the ith line segment is given by

$$a_i = \frac{x_{(i)} - x_{(i-1)}}{i/n - (i-1)/n} = \frac{x_{(i)} - x_{(i-1)}}{1/n}$$

The inverse cdf is calculated by

$$X = \hat{F}^{-1}(R) = x_{(i-1)} + a_i\left(R - \frac{(i-1)}{n}\right) \tag{8.10}$$

when $(i-1)/n < R \leq i/n$.

For example, if a random number $R_1 = 0.71$ is generated, then R_1 is seen to lie in the fourth interval (between $3/5 = 0.60$ and $4/5 = 0.80$); so, by Equation (8.10),

$$\begin{aligned} X_1 &= x_{(4-1)} + a_4(R_1 - (4-1)/n) \\ &= 1.45 + 1.90(0.71 - 0.60) \\ &= 1.66 \end{aligned}$$

The reader is referred to Figure 8.4 for a graphical view of the generation procedure.

In Example 8.2, each data point was represented in the empirical cdf. If a large sample of data is available (and sample sizes from several hundred to tens of thousands are possible with modern, automated data collection), then it might be more convenient (and computationally efficient) to first summarize the data into a frequency distribution with a much smaller number of intervals and then fit a continuous empirical cdf to the frequency distribution. Only a slight generalization of Equation (8.10) is required to accomplish this. Now the slope of the ith line segment is given by:

$$a_i = \frac{x_{(i)} - x_{(i-1)}}{c_i - c_{i-1}}$$

where c_i is the cumulative probability of the first i intervals of the frequency distribution and $x_{(i-1)} < x \leq x_{(i)}$ is the ith interval. The inverse cdf is calculated as

$$X = \hat{F}^{-1}(R) = x_{(i-1)} + a_i(R - c_{i-1}) \tag{8.11}$$

when $c_{i-1} < R \leq c_i$.

Example 8.3

Suppose that 100 broken-widget repair times have been collected. The data are summarized in Table 8.3 in terms of the number of observations in various intervals. For example, there were 31 observations between 0 and 0.5 hour, 10 between 0.5 and 1 hour, and so on. Suppose it is known that all repairs take at least 15 minutes, so that $X \geq 0.25$ hour always. Then we set $x_{(0)} = 0.25$, as shown in Table 8.3 and Figure 8.5.

For example, suppose the first random number generated is $R_1 = 0.83$. Then, since R_1 is between $c_3 = 0.66$ and $c_4 = 1.00$,

$$X_1 = x_{(4-1)} + a_4(R_1 - c_{4-1}) = 1.5 + 1.47(0.83 - 0.66) = 1.75 \tag{8.12}$$

Table 8.3 Summary of Repair-Time Data

i	*Interval (Hours)*	*Frequency*	*Relative Frequency*	*Cumulative Frequency, c_i*	*Slope a_i*
1	$0.25 \le x \le 0.5$	31	0.31	0.31	0.81
2	$0.5 < x \le 1.0$	10	0.10	0.41	5.0
3	$1.0 < x \le 1.5$	25	0.25	0.66	2.0
4	$1.5 < x \le 2.0$	34	0.34	1.00	1.47

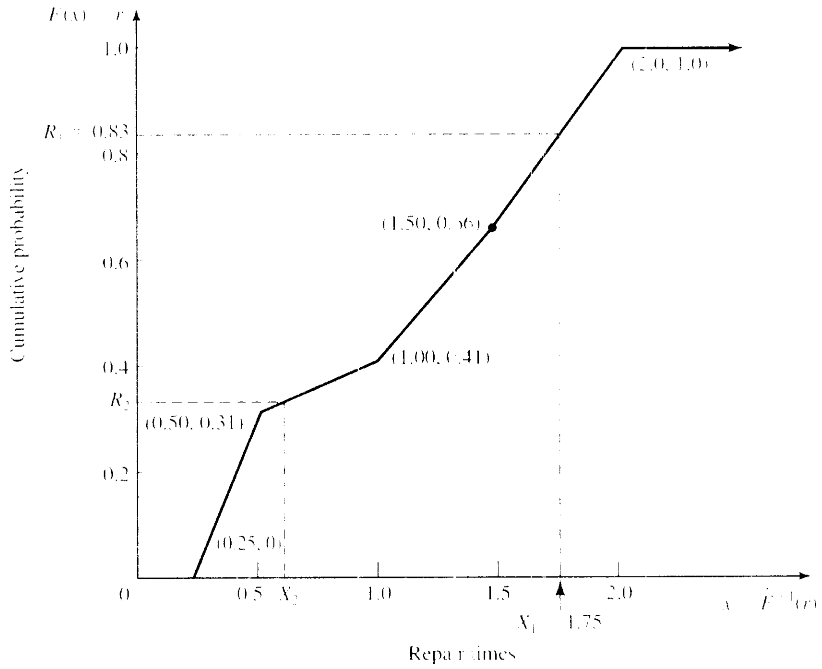

Figure 8.5 Generating variates from the empirical distribution function for repair-time data ($X \ge 0.25$).

As another illustration, suppose that $R_2 = 0.33$. Since $c_1 = 0.31 < R_2 \le 0.41 = c_2$,

$$\begin{aligned} X_2 &= x_{(1)} + a_2(R_2 - c_1) \\ &= 0.5 + 5.0(0.33 - 0.31) \\ &= 0.6 \end{aligned}$$

The point ($R_2 = 0.33$, $X_2 = 0.6$) is also shown in Figure 8.5.

Now reconsider the data of Table 8.3. The data are restricted in the range $0.25 \le X \le 2.0$, but the underlying distribution might have a wider range. This provides one important reason for attempting to find a theoretical statistical distribution (such as the gamma or Weibull) for the data: that these distributions do allow a wider range—namely, $0 \le X < \infty$. On the other hand, an empirical distribution adheres closely to what is present in the data itself, and the data are often the best source of information available.

When data are summarized in terms of frequency intervals, it is recommended that relatively short intervals be used, for doing so results in a more accurate portrayal of the underlying cdf. For example, for the repair-time data of Table 8.3, for which there were $n = 100$ observations, a much more accurate estimate could have been obtained by using 10 to 20 intervals, certainly not an excessive number, rather than the four fairly wide intervals actually used here for purposes of illustration.

Several comments are in order:

1. A computerized version of the procedure will become more inefficient as the number of intervals, n, increases. A systematic computerized version is often called a table-lookup generation scheme, because, given a value of R, the computer program must search an array of c_i values to find the interval i in which R lies, namely the interval i satisfying

$$c_{i-1} < R \leq c_i$$

The more intervals there are, the longer on the average the search will take if it is implemented in the crude way described here. The analyst should consider this trade-off between accuracy of the estimating cdf and computational efficiency when programming the procedure. If a large number of observations are available, the analyst may well decide to group the observations into from 20 to 50 intervals (say) and then use the procedure of Example 8.3—or a more efficient table-lookup procedure could be used, such as the one described in Law and Kelton [2000].

2. In Example 8.2, it was assumed that response times X satisfied $0 \leq X \leq 2.76$. This assumption led to the inclusion of the points $x_{(0)} = 0$ and $x_{(5)} = 2.76$ in Figure 8.4 and Table 8.2. If it is known a priori that X falls in some other range, for example, if it is known that response times are always between 15 seconds and 3 minutes—that is,

$$0.25 \leq X \leq 3.0$$

—then the points $x_{(0)} = 0.25$ and $x_{(6)} = 3.0$ would be used to estimate the empirical cdf of response times. Notice that, because of inclusion of the new point $x_{(0)}$, there are now six intervals instead of five, and each interval is assigned probability $1/6 = 0.167$.

8.1.6 Continuous Distributions without a Closed-Form Inverse

A number of useful continuous distributions do not have a closed form expression for their cdf or its inverse; examples include the normal, gamma, and beta distributions. For this reason, it is often stated that the inverse-transform technique for random-variate generation is not available for these distributions. It can, in effect, become available if we are willing to *approximate* the inverse cdf, or numerically integrate and search the cdf. Although this approach sounds inaccurate, notice that even a closed-form inverse requires approximation in order to evaluate it on a computer. For example, generating exponentially distributed random variates via the inverse cdf $X = F^{-1}(R) = -\ln(1 - R)/\lambda$ requires a numerical approximation for the logarithm function. Thus, there is no essential difference between using an approximate inverse cdf and approximately evaluating a closed-form inverse. The problem with using approximate inverse cdfs is that some of them are computationally slow to evaluate.

To illustrate the idea, consider a simple approximation to the inverse cdf of the standard normal distribution, proposed by Schmeiser [1979]:

$$X = F^{-1}(R) \approx \frac{R^{0.135} - (1 - R)^{0.135}}{0.1975}$$

This approximation gives at least one-decimal-place accuracy for $0.0013499 \leq R \leq 0.9986501$. Table 8.4 compares the approximation with exact values (to four decimal places) obtained by numerical integration for several values of R. Much more accurate approximations exist that are only slightly more complicated. A good source of these approximations for a number of distributions is Bratley, Fox, and Schrage [1996].

Table 8.4 Comparison of Approximate Inverse with Exact Values (To Four Decimal Places) for the Standard Normal Distribution

R	*Approximate Inverse*	*Exact Inverse*
0.01	−2.3263	−2.3373
0.10	−1.2816	−1.2813
0.25	−0.6745	−0.6713
0.50	0.0000	0.0000
0.75	0.6745	0.6713
0.90	1.2816	1.2813
0.99	2.3263	2.3373

8.1.7 Discrete Distributions

All discrete distributions can be generated via the inverse-transform technique, either numerically through a table-lookup procedure or, in some cases, algebraically, the final generation scheme being in terms of a formula. Other techniques are sometimes used for certain distributions, such as the convolution technique for the binomial distribution. Some of these methods are discussed in later sections. This subsection gives examples covering both empirical distributions and two of the standard discrete distributions, the (discrete) uniform and the geometric. Highly efficient table-lookup procedures for these and other distributions are found in Bratley, Fox, and Schrage [1996] and in Ripley [1987].

Example 8.4: An Empirical Discrete Distribution

At the end of any day, the number of shipments on the loading dock of the IHW Company (whose main product is the famous "incredibly huge widget") is either 0, 1, or 2, with observed relative frequency of occurrence of 0.50, 0.30, and 0.20, respectively. Internal consultants have been asked to develop a model to improve the efficiency of the loading and hauling operations; as part of this model, they will need to be able to generate values, X, to represent the number of shipments on the loading dock at the end of each day. The consultants decide to model X as a discrete random variable with the distribution given in Table 8.5 and shown in Figure 8.6.

The probability mass function (pmf), $p(x)$, is given by

$$p(0) = P(X = 0) = 0.50$$
$$p(1) = P(X = 1) = 0.30$$
$$p(2) = P(X = 2) = 0.20$$

Table 8.5 Distribution of Number of Shipments, X

x	$p(x)$	$F(x)$
0	0.50	0.50
1	0.30	0.80
2	0.20	1.00

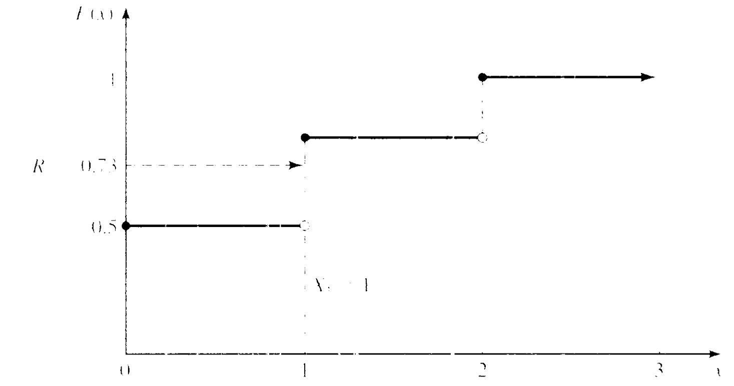

Figure 8.6 Cdf of number of shipments, X.

and the cdf, $F(x) = P(X \le x)$, is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.5 & 0 \le x < 1 \\ 0.8 & 1 \le x < 2 \\ 1.0 & 2 \le x \end{cases}$$

Recall that the cdf of a discrete random variable always consists of horizontal line segments with jumps of size $p(x)$ at those points, x, that the random variable can assume. For example, in Figure 8.6, there is a jump of size $p(0) = 0.5$ at $x = 0$, of size $p(1) = 0.3$ at $x = 1$, and of size $p(2) = 0.2$ at $x = 2$.

For generating discrete random variables, the inverse transform technique becomes a table-lookup procedure, but, unlike in the case of continuous variables, interpolation is not required. To illustrate the procedure, suppose that $R_1 = 0.73$ is generated. Graphically, as illustrated in Figure 8.6, first locate $R_1 = 0.73$ on the vertical axis, next draw a horizontal line segment until it hits a "jump" in the cdf, and then drop a perpendicular to the horizontal axis to get the generated variate. Here $R_1 = 0.73$ is transformed to $X_1 = 1$. This procedure is analogous to the procedure used for empirical continuous distributions in Section 8.1.5 and illustrated in Figure 8.4, except that the final step, linear interpolation, is eliminated.

The table-lookup procedure is facilitated by construction of a table such as Table 8.6. When $R_1 = 0.73$ is generated, first find the interval in which R_1 lies. In general, for $R = R_1$, if

$$F(x_{i-1}) = r_{i-1} < R \le r_i = F(x_i) \tag{8.13}$$

then set $X_1 = x_i$. Here, $r_0 = 0$; $x_0 = -\infty$; $x_1, x_2, \ldots, x_n$ are the possible values of the random variable; and $r_k = p(x_1) + \cdots + p(x_k)$, $k = 1, 2, \ldots, n$. For this example, $n = 3$, $x_1 = 0$, $x_2 = 1$, and $x_3 = 2$; hence, $r_1 = 0.5$, $r_2 = 0.8$, and $r_3 = 1.0$. (Notice that $r_n = 1.0$ in all cases.)

Table 8.6 Table for Generating the Discrete Variate X

i	*Input* r_i	*Output* x_i
1	0.50	0
2	0.80	1
3	1.00	2

Since $r_1 = 0.5 < R_1 = 0.73 \le r_2 = 0.8$, set $X_1 = x_2 = 1$. The generation scheme is summarized as follows:

$$X = \begin{cases} 0, & R \le 0.5 \\ 1, & 0.5 < R \le 0.8 \\ 2, & 0.8 < R \le 1.0 \end{cases}$$

Example 8.4 illustrated the table-lookup procedure; the next example illustrates an algebraic approach that can be used for certain distributions.

Example 8.5: A Discrete Uniform Distribution

Consider the discrete uniform distribution on $\{1, 2, \ldots, k\}$ with pmf and cdf given by

$$p(x) = \frac{1}{k}, \quad x = 1, 2, \ldots, k$$

and

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{k}, & 1 \le x < 2 \\ \frac{2}{k}, & 2 \le x < 3 \\ \vdots & \vdots \\ \frac{k-1}{k}, & k-1 \le x < k \\ 1, & k \le x \end{cases}$$

Let $x_i = i$ and $r_i = p(1) + \cdots + p(x_i) = F(x_i) = i/k$ for $i = 1, 2, \ldots, k$. Then, from Inequality (8.13), it can be seen that, if the generated random number R satisfies

$$r_{i-1} = \frac{i-1}{k} < R \le r_i = \frac{i}{k} \qquad (8.14)$$

then X is generated by setting $X = i$. Now Inequality (8.14) can be solved for i:

$$\begin{aligned} i - 1 &< Rk \le i \\ Rk &\le i < Rk + 1 \end{aligned} \qquad (8.15)$$

Let $\lceil y \rceil$ denote the smallest integer $\ge y$. For example, $\lceil 7.82 \rceil = 8$, $\lceil 5.13 \rceil = 6$, and $\lceil -1.32 \rceil = -1$. For $y \ge 0$, $\lceil y \rceil$ is a function that rounds up. This notation and Inequality (8.15) yield a formula for generating X, namely

$$X = \lceil Rk \rceil \qquad (8.16)$$

For example, consider the generating of a random variate X that is uniformly distributed on $\{1, 2, \ldots, 10\}$. The variate, X, might represent the number of pallets to be loaded onto a truck. Using Table A.1 as a source of random numbers R and using Equation (8.16) with $k = 10$ yields

$$
\begin{aligned}
R_1 &= 0.78 & X_1 &= \lceil 7.8 \rceil = 8\\
R_2 &= 0.03 & X_2 &= \lceil 0.3 \rceil = 1\\
R_3 &= 0.23 & X_3 &= \lceil 2.3 \rceil = 3\\
R_4 &= 0.97 & X_4 &= \lceil 9.7 \rceil = 10
\end{aligned}
$$

The procedure discussed here can be modified to generate a discrete uniform random variate with any range consisting of consecutive integers. Exercise 13 asks the student to devise a procedure for one such case.

Example 8.6: The Geometric Distribution

Consider the geometric distribution with pmf

$$p(x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots$$

where $0 < p < 1$. Its cdf is given by

$$
\begin{aligned}
F(x) &= \sum_{j=0}^{x} p(1-p)^j\\
&= \frac{p\{1-(1-p)^{x+1}\}}{1-(1-p)}\\
&= 1-(1-p)^{x+1}
\end{aligned}
$$

for $x = 0, 1, 2, \ldots$ Using the inverse-transform technique [i.e., Inequality (8.13)], recall that a geometric random variable X will assume the value x whenever

$$F(x-1) = 1-(1-p)^x < R \leq 1-(1-p)^{x+1} = F(x) \tag{8.17}$$

where R is a generated random number assumed $0 < R < 1$. Solving Inequality (8.17) for x proceeds as follows:

$$
\begin{aligned}
(1-p)^{x+1} &\leq 1-R < (1-p)^x\\
(x+1)\ln(1-p) &\leq \ln(1-R) < x\ln(1-p)
\end{aligned}
$$

But $1-p < 1$ implies that $\ln(1-p) < 0$, so that

$$\frac{\ln(1-R)}{\ln(1-p)} - 1 \leq x < \frac{\ln(1-R)}{\ln(1-p)} \tag{8.18}$$

Thus, $X = x$ for that integer value of x satisfying Inequality (8.18). In brief, and using the round-up function $\lceil \cdot \rceil$, we have

$$X = \left\lceil \frac{\ln(1-R)}{\ln(1-p)} - 1 \right\rceil \tag{8.19}$$

Since p is a fixed parameter, let $\beta = -1/\ln(1-p)$. Then $\beta > 0$ and, by Equation (8.19), $X = \lceil -\beta \ln(1-R) - 1 \rceil$. By Equation (8.1), $-\beta \ln(1-R)$ is an exponentially distributed random variable with mean β; so one way of generating a geometric variate with parameter p is to generate (by any method) an exponential variate with parameter $\beta^{-1} = -\ln(1-p)$, subtract one, and round up.

Occasionally, there is needed a geometric variate X that can assume values $\{q, q+1, q+2, \ldots\}$ with pmf $p(x) = p(1-p)^{x-q}(x = q, q+1, \ldots)$. Such a variate X can be generated, via Equation (8.19), by

$$X = q + \left\lceil \frac{\ln(1-R)}{\ln(1-p)} - 1 \right\rceil \tag{8.20}$$

One of the most common cases is $q = 1$.

Example 8.7

Generate three values from a geometric distribution on the range $\{X \geq 1\}$ with mean 2. Such a geometric distribution has pmf $p(x) = p(1-p)^{x-1}(x = 1, 2, \ldots)$ with mean $1/p = 2$, or $p = 1/2$. Thus, X can be generated by Equation (8.20) with $q = 1$, $p = 1/2$, and $1/\ln(1-p) = -1.443$. Using Table A.1, $R_1 = 0.932$, $R_2 = 0.105$, and $R_3 = 0.687$ yields

$$\begin{aligned} X_1 &= 1 + \lceil -1.443 \ln(1-0.932) - 1 \rceil \\ &= 1 + \lceil 3.878 - 1 \rceil = 4 \\ X_2 &= 1 + \lceil -1.443 \ln(1-0.105) - 1 \rceil = 1 \\ X_3 &= 1 + \lceil -1.443 \ln(1-0.687) - 1 \rceil = 2 \end{aligned}$$

Exercise 15 deals with an application of the geometric distribution

8.2 ACCEPTANCE–REJECTION TECHNIQUE

Suppose that an analyst needed to devise a method for generating random variates, X, uniformly distributed between 1/4 and 1. One way to proceed would be to follow these steps:

Step 1. Generate a random number R.

Step 2a. If $R \geq 1/4$, accept $X = R$, then go to Step 3.

Step 2b. If $R < 1/4$, reject R, and return to Step 1.

Step 3. If another uniform random variate on $[1/4, 1]$ is needed, repeat the procedure beginning at Step 1. If not, stop.

Each time Step 1 is executed, a new random number R must be generated. Step 2a is an "acceptance" and Step 2b is a "rejection" in this acceptance–rejection technique. To summarize the technique, random variates (R) with some distribution (here uniform on $[0, 1]$) are generated until some condition ($R > 1/4$) is satisfied. When the condition is finally satisfied, the desired random variate, X (here uniform on $[1/4, 1]$), can be computed ($X = R$). This procedure can be shown to be correct by recognizing that the accepted values of R are conditioned values; that is, R itself does not have the desired distribution, but R conditioned on the event $\{R \geq 1/4\}$ does have the desired distribution. To show this, take $1/4 \leq a < b \leq 1$; then

$$P(a < R \leq b \mid 1/4 \leq R \leq 1) = \frac{P(a < R \leq b)}{P(1/4 \leq R \leq 1)} = \frac{b-a}{3/4} \tag{8.21}$$

which is the correct probability for a uniform distribution on $[1/4, 1]$. Equation (8.21) says that the probability distribution of R, given that R is between 1/4 and 1 (all other values of R are thrown out), is the desired distribution. Therefore, if $1/4 \leq R \leq 1$, set $X = R$.

The efficiency of an acceptance–rejection technique depends heavily on being able to minimize the number of rejections. In this example, the probability of a rejection is $P(R < 1/4) = 1/4$, so that the number of rejections is a geometrically distributed random variable with probability of "success" being $p = 3/4$ and mean number of rejections $(1/p - 1) = 4/3 - 1 = 1/3$. (Example 8.6 discussed the geometric distribution.) The mean number of random numbers R required to generate one variate X is one more than the number of rejections; hence, it is $4/3 = 1.33$. In other words, to generate 1000 values of X would require approximately 1333 random numbers R.

In the present situation, an alternative procedure exists for generating a uniform variate on [1/4, 1]—namely, Equation (8.5), which reduces to $X = 1/4 + (3/4)R$. Whether the acceptance–rejection technique or an alternative procedure, such as the inverse-transform technique [Equation (8.5)], is the more efficient depends on several considerations. The computer being used, the skills of the programmer and the relative inefficiency of generating the additional (rejected) random numbers needed by acceptance–rejection should be compared to the computations required by the alternative procedure. In practice, concern with generation efficiency is left to specialists who conduct extensive tests comparing alternative methods (i.e., until a simulation model begins to require excessive computer runtime due to the generator being used).

For the uniform distribution on [1/4, 1], the inverse-transform technique of Equation (8.5) is undoubtedly much easier to apply and more efficient than the acceptance–rejection technique. The main purpose of this example was to explain and motivate the basic concept of the acceptance–rejection technique. However, for some important distributions, such as the normal, gamma and beta, the inverse cdf does not exist in closed form and therefore the inverse-transform technique is difficult. These more advanced techniques are summarized by Bratley, Fox, and Schrage [1996], Fishman [1978], and Law and Kelton [2000].

In the following subsections, the acceptance–rejection technique is illustrated for the generation of random variates for the Poisson, nonstationary Poisson, and gamma distributions.

8.2.1 Poisson Distribution

A Poisson random variable, N, with mean $\alpha > 0$ has pmf

$$p(n) = P(N = n) = \frac{e^{-\alpha}\alpha^n}{n!}, \quad n = 0, 1, 2, \ldots$$

More important, however, is that N can be interpreted as the number of arrivals from a Poisson arrival process in one unit of time. Recall from Section 5.5 that the interarrival times, $A_1, A_2, \ldots$ of successive customers are exponentially distributed with rate α (i.e., α is the mean number of arrivals per unit time); in addition, an exponential variate can be generated by Equation (8.3). Thus, there is a relationship between the (discrete) Poisson distribution and the (continuous) exponential distribution:

$$N = n \tag{8.22}$$

if and only if

$$A_1 + A_2 + \cdots + A_n \le 1 < A_1 + \cdots + A_n + A_{n+1} \tag{8.23}$$

Equation (8.22), $N = n$, says there were exactly n arrivals during one unit of time; but Relation (23) says that the nth arrival occurred before time 1 while the $(n + 1)$st arrival occurred after time 1. Clearly, these two statements are equivalent. Proceed now by generating exponential interarrival times until some arrival, say $n + 1$, occurs after time 1; then set $N = n$.

For efficient generation purposes, Relation (8.23) is usually simplified by first using Equation (8.3), $A_i = (-1/\alpha)\ln R_i$, to obtain

$$\sum_{i=1}^{n} -\frac{1}{\alpha}\ln R_i \le 1 < \sum_{i=1}^{n+1} -\frac{1}{\alpha}\ln R_i$$

Next multiply through by $-\alpha$, which reverses the sign of the inequality, and use the fact that a sum of logarithms is the logarithm of a product, to get

$$\ln \prod_{i=1}^{n} R_i = \sum_{i=1}^{n} \ln R_i \geq -\alpha > \sum_{i=1}^{n+1} \ln R_i = \ln \prod_{i=1}^{n+1} R_i$$

Finally, use the relation $e^{\ln x} = x$ for any number x to obtain

$$\prod_{i=1}^{n} R_i \geq e^{-\alpha} > \prod_{i=1}^{n+1} R_i \tag{8.24}$$

which is equivalent to Relation (8.23). The procedure for generating a Poisson random variate, N, is given by the following steps:

Step 1. Set $n = 0$, $P = 1$.

Step 2. Generate a random number R_{n+1}, and replace P by $P \cdot R_{n+1}$.

Step 3. If $P < e^{-\alpha}$, then accept $N = n$. Otherwise, reject the current n, increase n by one, and return to step 2.

Notice that, upon completion of Step 2, P is equal to the rightmost expression in Relation (8.24). The basic idea of a rejection technique is again exhibited; if $P \geq e^{-\alpha}$ in Step 3, then n is rejected and the generation process must proceed through at least one more trial.

How many random numbers will be required, on the average, to generate one Poisson variate, N? If $N = n$, then $n + 1$ random numbers are required, so the average number is given by

$$E(N + 1) = \alpha + 1$$

which is quite large if the mean, α, of the Poisson distribution is large.

Example 8.8

Generate three Poisson variates with mean $\alpha = 0.2$. First, compute $e^{-\alpha} = e^{-0.2} = 0.8187$. Next, get a sequence of random numbers R from Table A.1 and follow the previously described Steps 1 to 3:

Step 1. Set $n = 0$, $P = 1$.

Step 2. $R_1 = 0.4357$, $P = 1 \cdot R_1 = 0.4357$.

Step 3. Since $P = 0.4357 < e^{-\alpha} = 0.8187$, accept $N = 0$.

Step 1–3. ($R_1 = 0.4146$ leads to $N = 0$.)

Step 1. Set $n = 0$, $P = 1$.

Step 2. $R_1 = 0.8353$, $P = 1 \cdot R_1 = 0.8353$.

Step 3. Since $P \geq e^{-\alpha}$, reject $n = 0$ and return to Step 2 with $n = 1$.

Step 2. $R_2 = 0.9952$, $P = R_1 R_2 = 0.8313$.

Step 3. Since $P \geq e^{-\alpha}$, reject $n = 1$ and return to Step 2 with $n = 2$.

Step 2. $R_3 = 0.8004$, $P = R_1 R_2 R_3 = 0.6654$.

Step 3. Since $P < e^{-\alpha}$, accept $N = 2$.

The calculations required for the generation of these three Poisson random variates are summarized as follows:

n	R_{n+1}	P	*Accept/Reject*	*Result*
0	0.4357	0.4357	$P < e^{-\alpha}$ (accept)	$N = 0$
0	0.4146	0.4146	$P < e^{-\alpha}$ (accept)	$N = 0$
0	0.8353	0.8353	$P \geq e^{-\alpha}$ (reject)	
1	0.9952	0.8313	$P \geq e^{-\alpha}$ (reject)	
2	0.8004	0.6654	$P < e^{-\alpha}$ (accept)	$N = 2$

It took five random numbers, R, to generate three Poisson variates here ($N = 0$, $N = 0$, and $N = 2$), but in the long run, to generate, say, 1000 Poisson variates with mean $\alpha = 0.2$, it would require approximately $1000(\alpha + 1)$ or 1200 random numbers.

Example 8.9

Buses arrive at the bus stop at Peachtree and North Avenue according to a Poisson process with a mean of one bus per 15 minutes. Generate a random variate, N, which represents the number of arriving buses during a 1-hour time slot. Now, N is Poisson distributed with a mean of four buses per hour. First compute $e^{-\alpha} = e^{-4} = 0.0183$. Using a sequence of 12 random numbers from Table A.1 yields the following summarized results:

n	R_{n+1}	P	*Accept/Reject*	*Result*
0	0.4357	0.4357	$P \geq e^{-\alpha}$ (reject)	
1	0.4146	0.1806	$P \geq e^{-\alpha}$ (reject)	
2	0.8353	0.1508	$P \geq e^{-\alpha}$ (reject)	
3	0.9952	0.1502	$P \geq e^{-\alpha}$ (reject)	
4	0.8004	0.1202	$P \geq e^{-\alpha}$ (reject)	
5	0.7945	0.0955	$P \geq e^{-\alpha}$ (reject)	
6	0.1530	0.0146	$P < e^{-\alpha}$ (accept)	$N = 6$

It is immediately seen that a larger value of α (here $\alpha = 4$) usually requires more random numbers; if 1000 Poisson variates were desired, approximately $1000(\alpha + 1) = 5000$ random numbers would be required.

When α is large, say $\alpha \geq 15$, the rejection technique outlined here becomes quite expensive, but fortunately an approximate technique based on the normal distribution works quite well. When the mean, α, is large, then

$$Z = \frac{N - \alpha}{\sqrt{\alpha}}$$

is approximately normally distributed with mean zero and variance 1; this observation suggests an approximate technique. First, generate a standard normal variate Z by Equation (8.28) in Section 8.3.1, then generate the desired Poisson variate, N, by

$$N = \lceil \alpha + \sqrt{\alpha} Z - 0.5 \rceil \tag{8.25}$$

where $\lceil \cdot \rceil$ is the round-up function described in Section 8.1.7. (If $\alpha + \sqrt{\alpha} Z - 0.5 < 0$, then set $N = 0$.) The "0.5" used in the formula makes the round-up function become a "round to the nearest integer" function. Equation (8.25) is not an acceptance–rejection technique, but, when used as an alternative to the acceptance–rejection method, it provides a fairly efficient and accurate method for generating Poisson variates with a large mean.

Table 8.7 Arrival Rate for NSPP Example

t (min)	*Mean Time between Arrivals (min)*	*Arrival Rate $\lambda(t)$ (arrivals/min)*
0	15	1/15
60	12	1/12
120	7	1/7
180	5	1/5
240	8	1/8
300	10	1/10
360	15	1/15
420	20	1/20
480	20	1/20

8.2.2 Nonstationary Poisson Process

Another type of acceptance–rejection method (which is also called "thinning") can be used to generate interarrival times from a nonstationary Poisson process (NSPP) with arrival rate $\lambda(t)$, $0 \le t \le T$. A NSPP is an arrival process with an arrival rate that varies with time; see Section 5.5.2.

Consider, for instance, the arrival-rate function given in Table 8.7 that changes every hour. The idea behind thinning is to generate a stationary Poisson arrival process at the fastest rate (1/5 customer per minute in the example), but "accept" or admit only a portion of the arrivals, thinning out just enough to get the desired time-varying rate. Next we give the generic algorithm, which generates T_i as the time of the ith arrival. Remember that, in a stationary Poisson arrival process, the times between arrivals are exponentially distributed.

Step 1. Let $\lambda^* = \max_{0 \le t \le T} \lambda(t)$ be the maximum of the arrival rate function and set $t = 0$ and $i = 1$.

Step 2. Generate E from the exponential distribution with rate λ^* and let $t = t + E$ (this is the arrival time of the stationary Poisson process).

Step 3. Generate random number R from the $U(0, 1)$ distribution. If $R \le \lambda(t)/\lambda^*$ then $T_i = t$ and $i = i + 1$.

Step 4. Go to Step 2.

The thinning algorithm can be inefficient if there are large differences between the typical and the maximum arrival rate. However, thinning has the advantage that it works for any integrable arrival rate function, not just a piecewise-constant function as in this example.

Example 8.10

For the arrival-rate function in Table 8.7, generate the first two arrival times.

Step 1. $\lambda^* = \max_{0 \le t \le T} \lambda(t) = 1/5$, $t = 0$ and $i = 1$.

Step 2. For random number $R = 0.2130$, $E = -5 \ln(0.213) = 13.13$ and $t = 0 + 13.13 = 13.13$.

Step 3. Generate $R = 0.8830$. Since $R = 0.8830 \not\le \lambda(13.13)/\lambda^* = (1/15)/(1/5) = 1/3$, do not generate the arrival.

Step 4. Go to Step 2.

Step 2. For random number $R = 0.5530$, $E = -5 \ln(0.553) = 2.96$, and $t = 13.13 + 2.96 = 16.09$.

Step 3. Generate $R = 0.0240$. Since $R = 0.0240 \le \lambda(16.09)/\lambda^* = (1/15)/(1/5) = 1/3$, set $T_1 = t = 16.09$ and $i = i + 1 = 2$.

Step 4. Go to Step 2.

Step 2. For random number $R = 0.0001$, $E = -5\ \ln(0.0001) = 46.05$ and $t = 16.09 + 46.05 = 62.14$.

Step 3. Generate $R = 0.1443$. Since $R = 0.1443 \le \lambda(62.14)/\lambda^* = (1/12)/(1/5) = 5/12$, set $T_i = t = 62.14$ and $i = i + 1 = 3$.

Step 4. Go to Step 2.

8.2.3 Gamma Distribution

Several acceptance–rejection techniques for generating gamma random variates have been developed. (See Bratley, Fox, and Schrage [1996]; Fishman [1978]; and Law and Kelton [2000].) One of the more efficient is by Cheng [1977]; the mean number of trials is between 1.13 and 1.47 for any value of the shape parameter $\beta \ge 1$.

If the shape parameter β is an integer, say $\beta = k$, one possibility is to use the convolution technique in Example 8.12, because the Erlang distribution is a special case of the more general gamma distribution. On the other hand, the acceptance–rejection technique described here would be a highly efficient method for the Erlang distribution especially if $\beta = k$ were large. The routine generates gamma random variates with scale parameter θ and shape parameter β—that is, with mean $1/\theta$ and variance $1/\beta\theta^2$. The steps are as follows:

Step 1. Compute $a = 1/(2\beta - 1)^{1/2}$, $b = \beta - \ln 4$.

Step 2. Generate R_1 and R_2. Set $V = R_1/(1-R_1)$.

Step 3. Compute $X = \beta V^a$.

Step 4a. If $X > b + (\beta a + 1)\ln(V) - \ln(R_1^2 R_2)$, reject X and return to Step 2.

Step 4b. If $X \le b + (\beta a + 1)\ln(V) - \ln(R_1^2 R_2)$, use X as the desired variate.

The generated variates from Step 4b will have mean and variance both equal to β. If it is desired to have mean $1/\theta$ and variance $1/\beta\theta^2$ as in Section 5.4, then include Step 5.

(**Step 5.** Replace X by $X/(\beta\theta)$.)

The basic idea of all acceptance–rejection methods is again illustrated here, but the proof of this example is beyond the scope of this book. In Step 3, $X = \beta V^a = \beta[R_1/(1-R_1)]^a$ is not gamma distributed, but rejection of certain values of X in Step 4a guarantees that the accepted values in Step 4b do have the gamma distribution.

Example 8.11

Downtimes for a high-production candy-making machine have been found to be gamma distributed with mean 2.2 minutes and variance 2.10 minutes². Thus, $1/\theta = 2.2$ and $1/\beta\theta^2 = 2.10$, which together imply that $\beta = 2.30$ and $\theta = 0.4545$.

Step 1. $a = 0.53$, $b = 0.91$.

Step 2. Generate $R_1 = 0.832$, $R_2 = 0.021$. Set $V = 0.832/(1 - 0.832) = 4.952$.

Step 3. Compute $X = 2.3(4.952)^{0.53} = 5.37$.

Step 4. $X = 5.37 > 0.91 + [2.3(0.53) + 1]\ \ln(4.952) - \ln[(0.832)^2 0.021] = 8.68$, so reject X and return to Step 2.

Step 2. Generate $R_1 = 0.434$, $R_2 = 0.716$. Set $V = 0.434/(1 - 0.434) = 0.767$.

Step 3. Compute $X = 2.3(0.767)^{0.53} = 2.00$.

Step 4. Since $X = 2.00 \le 0.91 + [2.3(0.53) + 1]\ \ln(0.767) - \ln[(0.434)^2 0.716] = 2.32$, accept X.

Step 5. Divide X by $\beta\theta = 1.045$ to get $X = 1.91$

This example took two trials (i.e., one rejection) to generate an acceptable gamma-distributed random variate, but, on the average, to generate, say, 1000 gamma variates, the method will require between 1130 and 1470 trials, or equivalently, between 2260 and 2940 random numbers. The method is somewhat cumbersome for hand calculations, but is easy to program on the computer and is one of the most efficient gamma generators known.

8.3 SPECIAL PROPERTIES

"Special properties" are just as the name implies. They are variate-generation techniques that are based on features of a particular family of probability distributions, rather than being general-purpose techniques like the inverse-transform or acceptance–rejection techniques.

8.3.1 Direct Transformation for the Normal and Lognormal Distributions

Many methods have been developed for generating normally distributed random variates. The inverse-transform technique cannot easily be applied, however, because the inverse cdf cannot be written in closed form. The standard normal cdf is given by

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt, \qquad -\infty < x < \infty$$

This section describes an intuitively appealing direct transformation that produces an independent pair of standard normal variates with mean zero and variance 1. The method is due to Box and Muller [1958]. Although not as efficient as many more modern techniques, it is easy to program in a scientific language, such as FORTRAN, C, C++, Visual Basic, or Java. We then show how to transform a standard normal variate into a normal variate with mean μ and variance σ^2. Once we have a method (this or any other) for generating X from a $N(\mu, \sigma^2)$ distribution, then we can generate a lognormal random variate Y with parameters μ and σ^2 by using the direct transformation $Y = e^X$. (Recall that μ and σ^2 are *not* the mean and variance of the lognormal; see Equations (5.58) and (5.59).)

Consider two standard normal random variables, Z_1 and Z_2, plotted as a point in the plane as shown in Figure 8.7 and represented in polar coordinates as

$$\begin{aligned} Z_1 &= B\cos\theta \\ Z_2 &= B\sin\theta \end{aligned} \tag{8.26}$$

It is known that $B^2 = Z_1^2 + Z_2^2$ has the chi-square distribution with 2 degrees of freedom, which is equivalent to an exponential distribution with mean 2. Thus, the radius, B, can be generated by use of Equation (8.3):

$$B = (-2\ \ln R)^{1/2} \tag{8.27}$$

By the symmetry of the normal distribution, it seems reasonable to suppose, and indeed it is the case, that the angle is uniformly distributed between 0 and 2π radians. In addition, the radius, B, and the angle, θ, are mutually independent. Combining Equations (8.26) and (8.27) gives a direct method for generating two independent standard normal variates, Z_1 and Z_2, from two independent random numbers, R_1 and R_2:

$$\begin{aligned} Z_1 &= (-2\ \ln R_1)^{1/2}\cos(2\pi R_2) \\ Z_2 &= (-2\ \ln R_1)^{1/2}\sin(2\pi R_2) \end{aligned} \tag{8.28}$$

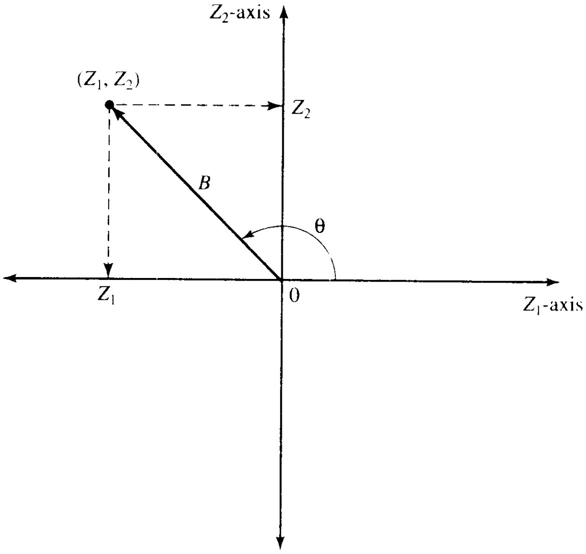

Figure 8.7 Polar representation of a pair of standard normal variables.

To illustrate the generation scheme, consider Equation (8.28) with $R_1 = 0.1758$ and $R_2 = 0.1489$. Two standard normal random variates are generated as follows:

$$Z_1 = [-2\ln(0.1758)]^{1/2}\cos(2\pi 0.1489) = 1.11$$
$$Z_2 = [-2\ln(0.1758)]^{1/2}\sin(2\pi 0.1489) = 1.50$$

To obtain normal variates X_i with mean μ and variance σ^2, we then apply the transformation

$$X_i = \mu + \sigma Z_i \tag{8.29}$$

to the standard normal variates. For example, to transform the two standard normal variates into normal variates with mean $\mu = 10$ and variance $\sigma^2 = 4$, we compute

$$X_1 = 10 + 2(1.11) = 12.22$$
$$X_2 = 10 + 2(1.50) = 13.00$$

8.3.2 Convolution Method

The probability distribution of a sum of two or more independent random variables is called a convolution of the distributions of the original variables. The convolution method thus refers to adding together two or more random variables to obtain a new random variable with the desired distribution. This technique can be applied to obtain Erlang variates and binomial variates. What is important is not the cdf of the desired random variable, but rather its relation to other variates more easily generated.

Example 8.12: Erlang Distribution

As was discussed in Section 5.4, an Erlang random variable X with parameters (k, θ) can be shown to be the sum of k independent exponential random variables, X_i, $i = 1, \ldots, k$, each having mean $1/k\theta$—that is,

$$X = \sum_{i=1}^{k} X_i$$

The convolution approach is to generate $X_1, X_2, \dots, X^k$, then sum them to get X. In the case of the Erlang, each X_i can be generated by Equation (8.3) with $1/\lambda = 1/k\theta$. Therefore, an Erlang variate can be generated by

$$\begin{aligned} X &= \sum_{i=1}^{k} -\frac{1}{k\theta} \ln R_i \\ &= -\frac{1}{k\theta} \ln\left(\prod_{i=1}^{k} R_i\right) \end{aligned} \tag{8.30}$$

It is more efficient computationally to multiply all the random numbers first and then to compute only one logarithm.

Example 8.13

Trucks arrive at a large warehouse in a completely random fashion that is modeled as a Poisson process with arrival rate $\lambda = 10$ trucks per hour. The guard at the entrance sends trucks alternately to the north and south docks. An analyst has developed a model to study the loading/unloading process at the south docks and needs a model of the arrival process at the south docks alone. An interarrival time X between successive truck arrivals at the south docks is equal to the sum of two interarrival times at the entrance and thus it is the sum of two exponential random variables, each having mean 0.1 hour, or 6 minutes. Thus, X has the Erlang distribution with $K = 2$ and mean $1/\theta = 2/\lambda = 0.2$ hour. To generate the variate X, first obtain $K = 2$ random numbers from Table A.1, say $R_1 = 0.937$ and $R_2 = 0.217$. Then, by Equation (8.30),

$$\begin{aligned} X &= 0.1 \ln[0.937(0.217)] \\ &= 0.159 \text{ hour } = 9.56 \text{ minutes} \end{aligned}$$

In general, Equation (8.30) implies that K uniform random number are needed for each Erlang variate generated. If K is large, it is more efficient to generate Erlang variates by other techniques, such as one of the many acceptance–rejection techniques for the gamma distribution given in Section 8.2.3, or by Bratley, Fox and Schrage [1996], Fishman [1978], and Law and Kelton [2000].

8.3.3 More Special Properties

There are many relationships among probability distributions that can be exploited for random-variate generation. The convolution method in the Section 8.3.2 is one example. Another particularly useful example is the relationship between the beta distribution and the gamma distribution.

Suppose that X_1 has a gamma distribution with shape parameter β_1 and scale parameter $\theta_1 = 1/\beta_1$, while X_2 has a gamma distribution with shape parameter β_2 and scale parameter $\theta_2 = 1/\beta_2$, and that these two random variables are independent. Then

$$Y = \frac{X_1}{X_1 + X_2}$$

has a beta distribution with parameters β_1 and β_2 on the interval (0, 1). If, instead, we want Y to be defined on the interval (a, b), then set

$$Y = a + (b - a)\left(\frac{X_1}{X_1 + X_2}\right)$$

Thus, using the acceptance–rejection technique for gamma variates defined in the previous section, we can generate beta variates, with two gamma variates required for each beta.

Although this method of beta generation is convenient, there are faster methods based on acceptance–rejection ideas. See, for instance, Devroye [1986] or Dagpunar [1988].

8.4 SUMMARY

The basic principles of random-variate generation via the inverse-transform technique, the acceptance–rejection technique, and special properties have been introduced and illustrated by examples. Methods for generating many of the important continuous and discrete distributions, plus all empirical distributions, have been given. See Schmeiser [1980] for an excellent survey; for a state-of-the-art treatment, the reader is referred to Devroye [1986] or Dagpunar [1988].

REFERENCES

BRATLEY, P., B. L. FOX, AND L. E. SCHRAGE [1996], *A Guide to Simulation*, 2d ed., Springer-Verlag, New York.

BOX, G. E. P., AND M. F. MULLER [1958], "A Note on the Generation of Random Normal Deviates," *Annals of Mathematical Statistics*, Vol. 29, pp. 610–611.

CHENG, R. C. H. [1977], "The Generation of Gamma Variables," *Applied Statistician*, Vol. 26, No. 1, pp. 71–75.

DAGPUNAR, J. [1988], *Principles of Random Variate Generation*, Clarendon Press, Oxford, New York.

DEVROYE, L. [1986], *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

FISHMAN, G. S. [1978], *Principles of Discrete Event Simulation*, Wiley, New York.

LAW, A. M., AND W. D. KELTON [2000], *Simulation Modeling & Analysis*, 3d ed., McGraw–Hill, New York.

RIPLEY, B. D. [1987], *Stochastic Simulation*, Wiley, New York.

SCHMEISER, BRUCE W. [1979], "Approximations to the Inverse Cumulative Normal Function for Use on Hand Calculators," *Applied Statistics*, Vol. 28, pp. 175–176.

SCHMEISER, B. W. [1980], "Random Variate Generation: A Survey," in *Simulation with Discrete Models: A State of the Art View*, T. I. Ören, C. M. Shub, and P. F. Roth, eds., IEEE, New York.

EXERCISES

1. Develop a random-variate generator for X with pdf

$$f(x)=\begin{cases}\dfrac{3x^2}{2}, & -1\le x\le 1\\ 0, & \text{otherwise}\end{cases}$$

2. Develop a generation scheme for the triangular distribution with pdf

$$f(x)=\begin{cases}\dfrac{1}{2}(x-2), & 2\le x\le 3\\ \dfrac{1}{2}\left(2-\dfrac{x}{3}\right), & 3< x\le 6\\ 0, & \text{otherwise}\end{cases}$$

Generate 10 values of the random variate, compute the sample mean, and compare it to the true mean of the distribution.

3. Develop the triangular random-variate generator with range (0, 12) and mode 5.

4. Develop a generator for a triangular distribution with range (1, 10) and a mean of 4.

5. Given the following cdf for a continuous variable with range from −3 to 4, develop a generator for the variable.

$$F(x) = \begin{cases} 0, & x \le -3 \\ \dfrac{1}{2} + \dfrac{x}{6}, & -3 < x \le 0 \\ \dfrac{1}{2} + \dfrac{x^2}{32}, & 0 < x \le 4 \\ 1, & x > 4 \end{cases}$$

6. Given the cdf $F(x) = x^4/16$ on $0 \le x \le 2$, develop a generator for this distribution.

7. Given the pdf $f(x) = x^2/9$ on $0 \le x \le 3$, develop a generator for this distribution.

8. The pdf of a random variable is

$$f(x) = \begin{cases} \dfrac{1}{5}, & 0 < x \le 3 \\ \dfrac{1}{8}, & 3 < x \le 9 \\ 0, & \text{otherwise} \end{cases}$$

Develop the random-variate generator.

9. The cdf of a discrete random variable X is given by

$$F(x) = \frac{x(x+1)(2x+1)}{n(n+1)(2n+1)}, \quad x = 1, 2, \ldots, n$$

When $n = 4$, generate three values of X, using $R_1 = 0.83$, $R_2 = 0.24$, and $R_3 = 0.57$.

10. Times to failure for an automated production process have been found to be randomly distributed with a Weibull distribution with parameters $\beta = 2$ and $\alpha = 10$. Derive Equation (8.6), and then use it to generate five values from this Weibull distribution, using five random numbers taken from Table A.1.

11. The details of time taken by a mechanic to repair a breakdown are

Repair Time Range (Hours)	*Frequency*
1–2	15
2–3	12
3–4	14
4–5	25
5–6	32
6–7	14

Develop a lookup table and generate five repair times using random numbers.

12. In an inventory system, the lead time is found to follow uniform distribution with mean 10 days and half width 3 days. Generate five lead times.

13. For a preliminary version of a simulation model, the number of pallets, X, to be loaded onto a truck at a loading dock was assumed to be uniformly distributed between 8 and 24. Devise a method for generating X, assuming that the loads on successive trucks are independent. Use the technique of Example 8.5 for discrete uniform distributions. Finally, generate loads for 10 successive trucks by using four-digit random numbers.

14. Develop a method for generating values from a negative binomial distribution with parameters p and k, as described in Section 5.3. Generate 3 values when $p = 0.8$ and $k = 2$. [*Hint*: Think about the definition of the negative binomial as the number of Bernoulli trials until the kth success.]

15. The weekly demand, X, for a slow-moving item has been found to be approximated well by a geometric distribution on the range $\{0, 1, 2, \ldots\}$ with mean weekly demand of 2.5 items. Generate 10 values of X, demand per week, using random numbers from Table A.1. (*Hint*: For a geometric distribution on the range $\{q, q + 1, \ldots\}$ with parameter p, the mean is $1/p + q - 1$.)

16. In Exercise 15, suppose that the demand has been found to have a Poisson distribution with mean 2.5 items per week. Generate 10 values of X, demand per week, using random numbers from Table A.1. Discuss the differences between the geometric and the Poisson distributions.

17. Service time of a bank teller is found to follow normal with $\mu = 5$ minutes and $\sigma = 1$ minute. Generate five service times.

18. The time to attend a breakdown call is found to follow exponential with a mean of 2 hours. Generate exponential random variates representing the time to attend.

19. A machine is taken out of production either if it fails or after 5 hours, whichever comes first. By running similar machines until failure, it has been found that time to failure, X, has the Weibull distribution with $\alpha = 8$, $\beta = 0.75$, and $\nu = 0$ (refer to Sections 5.4 and 8.1.3). Thus, the time until the machine is taken out of production can be represented as $Y = \min(X, 5)$. Develop a step-by-step procedure for generating Y.

20. In an art gallery, the arrival of visitors follow Poisson with a mean of 4 per hour. Generate the arrivals for the next 1 hour.

21. Develop a technique for generating a binomial random variable, X, via the convolution technique. [*Hint*: X can be represented as the number of successes in n independent Bernoulli trials, each success having probability p. Thus, $X = \sum_{i=1}^{n} X_i$, where $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$.]

22. Develop an acceptance–rejection technique for generating a geometric random variable, X, with parameter p on the range $\{0, 1, 2, \ldots\}$. (*Hint*: X can be thought of as the number of trials before the first success occurs in a sequence of independent Bernoulli trials.)

23. Write a computer program to generate exponential random variates for a given mean value. Generate 1000 values and verify the variates generated using chi-square test.

24. Develop a computer program to generate binomial random variates with p = probability of success, n = number of trials, and x = random number between 0 and 1.

25. Write a computer program to generate 500 normal random variates of given μ and σ values and prepare a histogram.

26. Many spreadsheet, symbolic-calculation, and statistical-analysis programs have built-in routines for generating random variates from standard distributions. Try to find out what variate-generation methods

are used in one of these packages by looking at the documentation. Should you trust a variate generator if the method is not documented?

27. Suppose that, somehow, we have available a source of exponentially distributed random variates with mean 1. Write an algorithm to generate random variates with a triangular distribution by transforming the exponentially distributed random variates. [*Hint*: First transform to obtain uniformly distributed random variates.]

28. A study is conducted on the arrival of customers in a bus stop during the post lunch period. The system starts at 12.30 P.M. and the arrival rate per hour during different intervals of time are

Time	*Arrival Rate/Hour*
12.30–1.30 P.M.	20
1.30–2.30 P.M.	35
2.30–3.30 P.M.	60
3.30–4.30 P.M.	80

Generate arrivals from this NSPP.

29. Generate 10 values from a beta distribution on the interval [0, 1] with parameters $\beta_1 = 1.47$ and $\beta_2 = 2.16$. Next transform them to be on the interval [−10, 20].

Part IV

Analysis of Simulation Data

9

Input Modeling

Input models provide the driving force for a simulation model. In the simulation of a queueing system, typical input models are the distributions of time between arrivals and of service times. For an inventory-system simulation, input models include the distributions of demand and of lead time. For the simulation of a reliability system, the distribution of time to failure of a component is an example of an input model.

In the examples and exercises in Chapters 2 and 3, the appropriate distributions were specified for you. In real-world simulation applications, however, coming up with appropriate distributions for input data is a major task from the standpoint of time and resource requirements. Regardless of the sophistication of the analyst, faulty models of the inputs will lead to outputs whose interpretation could give rise to misleading recommendations.

There are four steps in the development of a useful model of input data:

1. Collect data from the real system of interest. This often requires a substantial time and resource commitment. Unfortunately, in some situations it is not possible to collect data (for example, when time is extremely limited, when the input process does not yet exist, or when laws or rules prohibit the collection of data). When data are not available, expert opinion and knowledge of the process must be used to make educated guesses.
2. Identify a probability distribution to represent the input process. When data are available, this step typically begins with the development of a frequency distribution, or histogram, of the data. Given the frequency distribution and a structural knowledge of the process, a family of distributions is chosen. Fortunately, as was described in Chapter 5, several well-known distributions often provide good approximations in practice.
3. Choose parameters that determine a specific instance of the distribution family. When data are available, these parameters may be estimated from the data.

4. Evaluate the chosen distribution and the associated parameters for goodness of fit. Goodness of fit may be evaluated informally, via graphical methods, or formally, via statistical tests. The chi-square and the Kolmogorov–Smirnov tests are standard goodness-of-fit tests. If not satisfied that the chosen distribution is a good approximation of the data, then the analyst returns to the second step, chooses a different family of distributions, and repeats the procedure. If several iterations of this procedure fail to yield a fit between an assumed distributional form and the collected data, the empirical form of the distribution may be used, as was described in Section 8.1.5.

Each of these steps is discussed in this chapter. Although software is now widely available to accomplish Steps 2, 3, and 4—including such stand-alone programs as ExpertFit® and Stat::Fit® and such integrated programs as Arena's Input Processor and @Risk's BestFit®—it is still important to understand what the software does, so that it can be used appropriately. Unfortunately, software is not as readily available for input modeling when there is a relationship between two or more variables of interest or when no data are available. These two topics are discussed toward the end of the chapter.

9.1 DATA COLLECTION

Problems are found at the end of each chapter, as exercises for the reader, in textbooks about mathematics, physics, chemistry, and other technical subjects. Years and years of working these problems could give the reader the impression that data are readily available. Nothing could be further from the truth. Data collection is one of the biggest tasks in solving a real problem. It is one of the most important and difficult problems in simulation. And, even when data are available, they have rarely been recorded in a form that is directly useful for simulation input modeling.

"GIGO," or "garbage-in–garbage-out," is a basic concept in computer science, and it applies equally in the area of discrete-system simulation. Even when the model structure is valid, if the input data are inaccurately collected, inappropriately analyzed, or not representative of the environment, the simulation output data will be misleading and possibly damaging or costly when used for policy or decision making.

Example 9.1: The Laundromat

As budding simulation students, the first two authors had assignments to simulate the operation of an ongoing system. One of these systems, which seemed to be a rather simple operation, was a self-service laundromat with 10 washing machines and six dryers.

However, the data-collection aspect of the problem rapidly became rather enormous. The interarrival-time distribution was not homogeneous; it changed by time of day and by day of week. The laundromat was open 7 days a week for 16 hours per day, or 112 hours per week. It would have been impossible to cover the operation of the laundromat with the limited resources available (two students who were also taking four other courses) and with a tight time constraint (the simulation was to be completed in a 4-week period). Additionally, the distribution of time between arrivals during one week might not have been followed during the next week. As a compromise, a sample of times was selected, and the interarrival-time distributions were classified according to arrival rate (perhaps inappropriately) as "high," "medium," and "low."

Service-time distributions also presented a difficult problem from many perspectives. The proportion of customers demanding the various service combinations had to be observed and recorded. The simplest case was the customer desiring one washer followed by one dryer. However, a customer might choose two washing machines followed by one dryer, one dryer only, and so on. The customers used numbered machines, and it was possible to follow the customers via that reference, rather than remembering them by personal characteristics. Because of the dependence between washer demand and dryer demand for an individual customer, it would have been inappropriate to treat the service times for washers and dryers separately as independent variables.

Some customers waited patiently for their clothes to complete the washing or drying cycle, and then they removed their clothes promptly. Others left the premises and returned after their clothes had finished their cycle on the machine being used. In a very busy period, the manager would remove a customer's clothes after the cycle and set them aside in a basket. It was decided that service termination would be measured as that point in time at which the machine was emptied of its contents.

Also, machines would break down from time to time. The length of the breakdown varied from a few moments, when the manager repaired the machine, to several days (a breakdown on Friday night, requiring a part not in the laundromat storeroom, would not be fixed until the following Monday). The short-term repair times were recorded by the student team. The long-term repair completion times were estimated by the manager. Breakdowns then became part of the simulation.

Many lessons can be learned from an actual experience at data collection. The first five exercises at the end of this chapter suggest some situations in which the student can gain such experience.

The following suggestions might enhance and facilitate data collection, although they are not all inclusive.

1. A useful expenditure of time is in planning. This could begin by a practice or preobserving session. Try to collect data while preobserving. Devise forms for this purpose. It is very likely that these forms will have to be modified several times before the actual data collection begins. Watch for unusual circumstances, and consider how they will be handled. When possible, videotape the system and extract the data later by viewing the tape. Planning is important, even if data will be collected automatically (e.g., via computer data collection), to ensure that the appropriate data are available. When data have already been collected by someone else, be sure to allow plenty of time for converting the data into a usable format.
2. Try to analyze the data as they are being collected. Figure out whether the data being collected are adequate to provide the distributions needed as input to the simulation. Find out whether any data being collected are useless to the simulation. There is no need to collect superfluous data.
3. Try to combine homogeneous data sets. Check data for homogeneity in successive time periods and during the same time period on successive days. For example, check for homogeneity of data from 2:00 P.M. to 3:00 P.M. and 3:00 P.M. to 4:00 P.M., and check to see whether the data are homogeneous for 2:00 P.M. to 3:00 P.M. on Thursday and Friday. When checking for homogeneity, an initial test is to see whether the means of the distributions (the average interarrival times, for example) are the same. The two-sample t test can be used for this purpose. A more thorough analysis would require a test of the equivalence of the distributions, perhaps via a quantile-quantile plot (described later).
4. Be aware of the possibility of data censoring, in which a quantity of interest is not observed in its entirety. This problem most often occurs when the analyst is interested in the time required to complete some process (for example, produce a part, treat a patient, or have a component fail), but the process begins prior to, or finishes after the completion of, the observation period. Censoring can result in especially long process times being left out of the data sample.
5. To discover whether there is a relationship between two variables, build a scatter diagram. Sometimes an eyeball scan of the scatter diagram will indicate whether there is a relationship between two variables of interest. Section 9.7 describes models for statistically dependent input data.
6. Consider the possibility that a sequence of observations that appear to be independent actually has autocorrelation. Autocorrelation can exist in successive time periods or for successive customers.

For example, the service time for the ith customer could be related to the service time for the $(i + n)$th customer. A brief introduction to autocorrelation was provided in Section 7.4.2, and some input models that account for autocorrelation are presented in Section 9.7.

7. Keep in mind the difference between input data and output or performance data, and be sure to collect input data. Input data typically represent the uncertain quantities that are largely beyond the control of the system and will not be altered by changes made to improve the system. Output data, on the other hand, represent the performance of the system when subjected to the inputs, performance that we might be trying to improve. In a queueing simulation, the customer arrival times are usually inputs, whereas the customer delay is an output. Performance data are useful for model validation, however—see Chapter 10.

Again, these are just a few suggestions. As a rule, data collection and analysis must be approached with great care.

9.2 IDENTIFYING THE DISTRIBUTION WITH DATA

In this section, we discuss methods for selecting families of input distributions when data are available. The specific distribution within a family is specified by estimating its parameters, as described in Section 9.3. Section 9.6 takes up the case in which data are unavailable.

9.2.1 Histograms

A frequency distribution or histogram is useful in identifying the shape of a distribution. A histogram is constructed as follows:

1. Divide the range of the data into intervals. (Intervals are usually of equal width; however, unequal widths may be used if the heights of the frequencies are adjusted.)
2. Label the horizontal axis to conform to the intervals selected.
3. Find the frequency of occurrences within each interval.
4. Label the vertical axis so that the total occurrences can be plotted for each interval.
5. Plot the frequencies on the vertical axis.

The number of class intervals depends on the number of observations and on the amount of scatter or dispersion in the data. Hines, Montgomery, Goldsman, and Borrow [2002] state that choosing the number of class intervals approximately equal to the square root of the sample size often works well in practice. If the intervals are too wide, the histogram will be coarse, or blocky, and its shape and other details will not show well. If the intervals are too narrow, the histogram will be ragged and will not smooth the data. Examples of ragged, coarse, and appropriate histograms of the same data are shown in Figure 9.1. Modern data-analysis software often allows the interval sizes to be changed easily and interactively until a good choice is found.

The histogram for continuous data corresponds to the probability density function of a theoretical distribution. If continuous, a line drawn through the center point of each class interval frequency should result in a shape like that of a pdf.

Histograms for discrete data, where there are a large number of data points, should have a cell for each value in the range of the data. However, if there are few data points, it could be necessary to combine adjacent

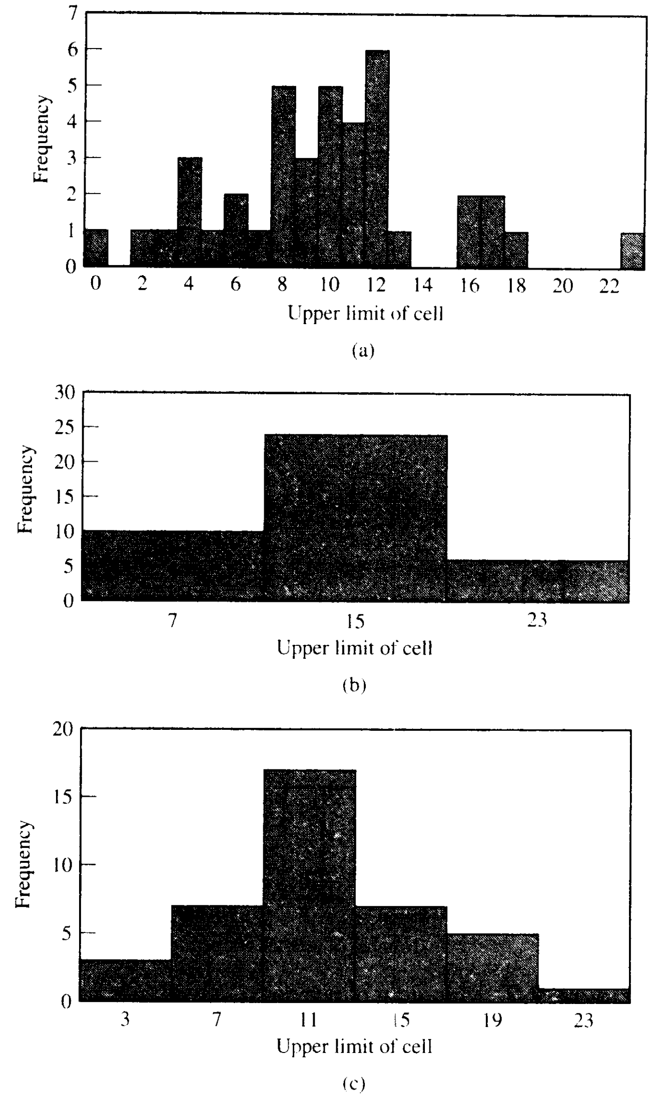

Figure 9.1 Ragged, coarse, and appropriate histograms: (a) original data—too ragged; (b) combining adjacent cells—too coarse; (c) combining adjacent cells—appropriate.

cells to eliminate the ragged appearance of the histogram. If the histogram is associated with discrete data, it should look like a probability mass function.

Example 9.2: Discrete Data

The number of vehicles arriving at the northwest corner of an intersection in a 5-minute period between 7:00 A.M. and 7:05 A.M. was monitored for five workdays over a 20-week period. Table 9.1 shows the resulting data. The first entry in the table indicates that there were 12 5-minute periods during which zero vehicles arrived, 10 periods during which one vehicle arrived, and so on.

The number of automobiles is a discrete variable, and there are ample data, so the histogram may have a cell for each possible value in the range of the data. The resulting histogram is shown in Figure 9.2.

Table 9.1 Number of Arrivals in a 5-Minute Period

Arrivals per Period	*Frequency*	*Arrivals per Period*	*Frequency*
0	12	6	7
1	10	7	5
2	19	8	5
3	17	9	3
4	10	10	3
5	8	11	1

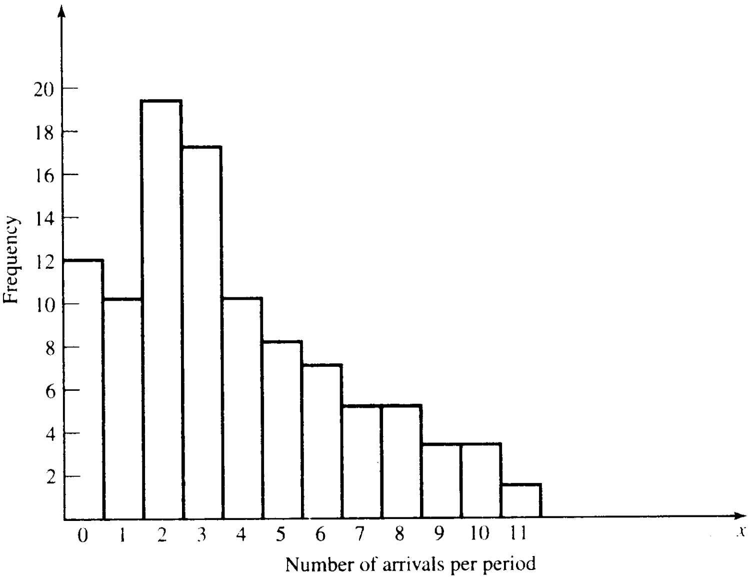

Figure 9.2 Histogram of number of arrivals per period.

Example 9.3: Continuous Data

Life tests were performed on a random sample of electronic components at 1.5 times the nominal voltage, and their lifetime (or time to failure), in days, was recorded:

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543
7.004	31.764	1.005	1.147	0.219
3.217	14.382	1.008	2.336	4.562

Table 9.2 Electronic Component Data

Component Life (Days)	*Frequency*
$0 \le x_j < 3$	23
$3 \le x_j < 6$	10
$6 \le x_j < 9$	5
$9 \le x_j < 12$	1
$12 \le x_j < 15$	1
$15 \le x_j < 18$	2
$18 \le x_j < 21$	0
$21 \le x_j < 24$	1
$24 \le x_j < 27$	1
$27 \le x_j < 30$	0
$30 \le x_j < 33$	1
$33 \le x_j < 36$	1
⋮	⋮
$42 \le x_j < 45$	1
⋮	⋮
$57 \le x_j < 60$	1
⋮	⋮
$78 \le x_j < 81$	1
⋮	⋮
$144 \le x_j < 147$	1

Lifetime, usually considered a continuous variable, is recorded here to three-decimal-place accuracy. The histogram is prepared by placing the data in class intervals. The range of the data is rather large, from 0.002 day to 144.695 days. However, most of the values (30 of 50) are in the zero-to-5-day range. Using intervals of width three results in Table 9.2. The data of Table 9.2 are then used to prepare the histogram shown in Figure 9.3.

9.2.2 Selecting the Family of Distributions

In Chapter 5, some distributions that arise often in simulation were described. Additionally, the shapes of these distributions were displayed. The purpose of preparing a histogram is to infer a known pdf or pmf. A family of distributions is selected on the basis of what might arise in the context being investigated along with the shape of the histogram. Thus, if interarrival-time data have been collected, and the histogram has a shape similar to the pdf in Figure 5.9, the assumption of an exponential distribution would be warranted. Similarly, if measurements of the weights of pallets of freight are being made, and the histogram appears

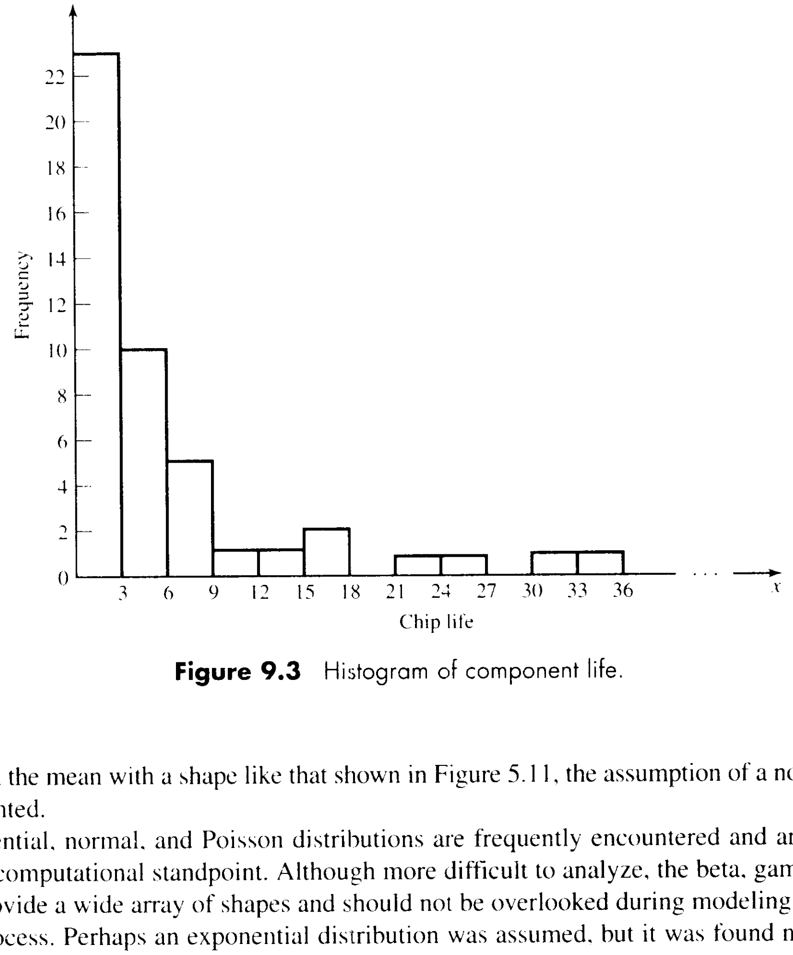

Figure 9.3 Histogram of component life.

symmetric about the mean with a shape like that shown in Figure 5.11, the assumption of a normal distribution would be warranted.

The exponential, normal, and Poisson distributions are frequently encountered and are not difficult to analyze from a computational standpoint. Although more difficult to analyze, the beta, gamma, and Weibull distributions provide a wide array of shapes and should not be overlooked during modeling of an underlying probabilistic process. Perhaps an exponential distribution was assumed, but it was found not to fit the data. The next step would be to examine where the lack of fit occurred. If the lack of fit was in one of the tails of the distribution, perhaps a gamma or Weibull distribution would fit the data more adequately.

There are literally hundreds of probability distributions that have been created; many were created with some specific physical process in mind. One aid to selecting distributions is to use the physical basis of the distributions as a guide. Here are some examples:

Binomial: Models the number of successes in n trials, when the trials are independent with common success probability, p; for example, the number of defective computer chips found in a lot of n chips.

Negative Binomial (includes the geometric distribution): Models the number of trials required to achieve k successes; for example, the number of computer chips that we must inspect to find 4 defective chips.

Poisson: Models the number of independent events that occur in a fixed amount of time or space; for example, the number of customers that arrive to a store during 1 hour, or the number of defects found in 30 square meters of sheet metal.

Normal: Models the distribution of a process that can be thought of as the sum of a number of component processes; for example, a time to assemble a product that is the sum of the times required for each assembly operation. Notice that the normal distribution admits negative values, which could be impossible for process times.

Lognormal: Models the distribution of a process that can be thought of as the product of (meaning to multiply together) a number of component processes—for example, the rate on an investment, when interest is compounded, is the product of the returns for a number of periods.

Exponential: Models the time between independent events, or a process time that is memoryless (knowing how much time has passed gives no information about how much additional time will pass before the process is complete)—for example, the times between the arrivals from a large population of potential customers who act independently of each other. The exponential is a highly variable distribution; it is sometimes overused, because it often leads to mathematically tractable models. Recall that, if the time between events is exponentially distributed, then the number of events in a fixed period of time is Poisson.

Gamma: An extremely flexible distribution used to model nonnegative random variables. The gamma can be shifted away from 0 by adding a constant.

Beta: An extremely flexible distribution used to model bounded (fixed upper and lower limits) random variables. The beta can be shifted away from 0 by adding a constant and can be given a range larger than [0, 1] by multiplying by a constant.

Erlang: Models processes that can be viewed as the sum of several exponentially distributed processes—for example, a computer network fails when a computer and two backup computers fail, and each has a time to failure that is exponentially distributed. The Erlang is a special case of the gamma.

Weibull: Models the time to failure for components—for example, the time to failure for a disk drive. The exponential is a special case of the Weibull.

Discrete or Continuous Uniform: Models complete uncertainty: All outcomes are equally likely. This distribution often is used inappropriately, when there are no data.

Triangular: Models a process for which only the minimum, most likely, and maximum values of the distribution are known; for example, the minimum, most likely, and maximum time required to test a product. This model is often a marked improvement over a uniform distribution.

Empirical: Resamples from the actual data collected; often used when no theoretical distribution seems appropriate.

Do not ignore physical characteristics of the process when selecting distributions. Is the process naturally discrete or continuous valued? Is it bounded, or is there no natural bound? This knowledge, which does not depend on data, can help narrow the family of distributions from which to choose. And keep in mind that there is no "true" distribution for any stochastic input process. An input model is an approximation of reality, so the goal is to obtain an approximation that yields useful results from the simulation experiment.

The reader is encouraged to complete Exercises 6 through 11 to learn more about the shapes of the distributions mentioned in this section. Examining the variations in shape as the parameters change is very instructive.

9.2.3 Quantile-Quantile Plots

The construction of histograms, as discussed in Section 9.2.1, and the recognition of a distributional shape, as discussed in Section 9.2.2, are necessary ingredients for selecting a family of distributions to represent a sample of data. However, a histogram is not as useful for evaluating the *fit* of the chosen distribution. When there is a small number of data points, say 30 or fewer, a histogram can be rather ragged. Further, our perception of the fit depends on the widths of the histogram intervals. But, even if the intervals are chosen well, grouping data into cells makes it difficult to compare a histogram to a continuous probability density function. A quantile–quantile ($q - q$) plot is a useful tool for evaluating distribution fit, one that does not suffer from these problems.

If X is a random variable with cdf F, then the q-quantile of X is that value γ such that $F(\gamma) = P(X \le \gamma) = q$, for $0 < q < 1$. When F has an inverse, we write $\gamma = F^{-1}(q)$.

Now let $\{x_i, i = 1, 2, \ldots, n\}$ be a sample of data from X. Order the observations from the smallest to the largest, and denote these as $\{y_j, j = 1, 2, \ldots, n\}$, where $y_1 \le y_2 \le \cdots \le y_n$. Let j denote the ranking or order number. Therefore, $j = 1$ for the smallest and $j = n$ for the largest. The $q - q$ plot is based on the fact that y_j is an estimate of the $(j - 1/2)/n$ quantile of X. In other words,

$$y_j \text{ is approximately } F^{-1}\left(\frac{j-\frac{1}{2}}{n}\right)$$

Now suppose that we have chosen a distribution with cdf F as a possible representation of the distribution of X. If F is a member of an appropriate family of distributions, then a plot of y_j versus $F^{-1}((j - 1/2)/n)$ will be *approximately a straight line.* If F is from an appropriate family of distributions and also has appropriate parameter values, then the line will have slope 1. On the other hand, if the assumed distribution is inappropriate, the points will deviate from a straight line, usually in a systematic manner. The decision about whether to reject some hypothesized model is subjective.

Example 9.4: Normal $Q - Q$ Plot

A robot is used to install the doors on automobiles along an assembly line. It was thought that the installation times followed a normal distribution. The robot is capable of measuring installation times accurately. A sample of 20 installation times was automatically taken by the robot, with the following results, where the values are in seconds:

99.79	99.56	100.17	100.33
100.26	100.41	99.98	99.83
100.23	100.27	100.02	100.47
99.55	99.62	99.65	99.82
99.96	99.90	100.06	99.85

The sample mean is 99.99 seconds, and the sample variance is $(0.2832)^2$ seconds^2. These values can serve as the parameter estimates for the mean and variance of the normal distribution. The observations are now ordered from smallest to largest as follows:

j	*Value*	*j*	*Value*	*j*	*Value*	*j*	Value
1	99.55	6	99.82	11	99.98	16	100.26
2	99.56	7	99.83	12	100.02	17	100.27
3	99.62	8	99.85	13	100.06	18	100.33
4	99.65	9	99.90	14	100.17	19	100.41
5	99.79	10	99.96	15	100.23	20	100.47

The ordered observations are then plotted versus $F^{-1}((j - 1/2)/20)$, for $j = 1, 2, \ldots, 20$, where F is the cdf of the normal distribution with mean 99.99 and variance $(0.2832)^2$, to obtain a $q - q$ plot. The plotted values are shown in Figure 9.4, along with a histogram of the data that has the density function of the normal distribution superimposed. Notice that it is difficult to tell whether the data are well represented by a normal

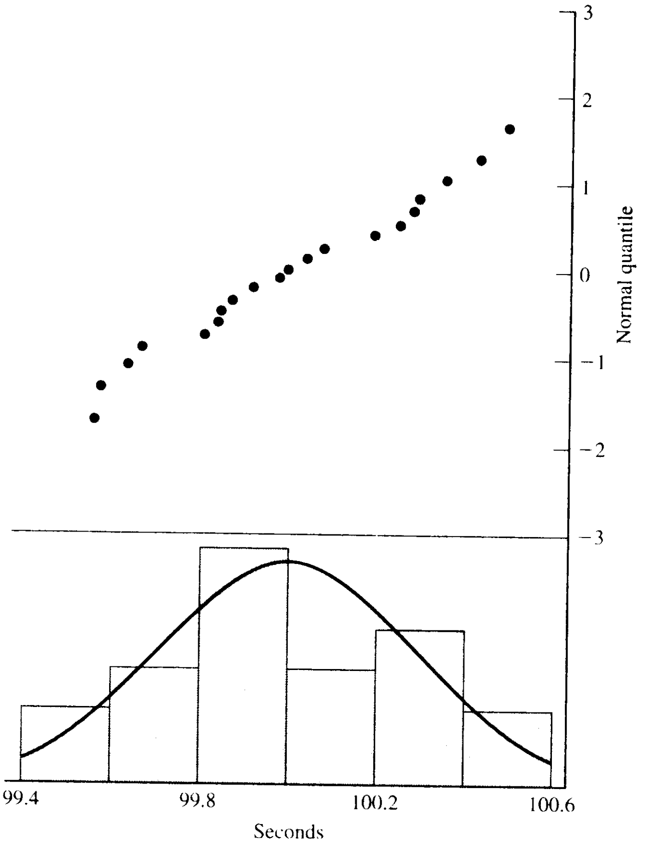

Figure 9.4 Histogram and $q - q$ plot of the installation times.

distribution from looking at the histogram, but the general perception of a straight line is quite clear in the $q - q$ plot and supports the hypothesis of a normal distribution.

In the evaluation of the linearity of a $q - q$ plot, the following should be considered:

1. The observed values will never fall exactly on a straight line.
2. The ordered values are not independent; they have been ranked. Hence, if one point is above a straight line, it is likely that the next point will also lie above the line. And it is unlikely that the points will be scattered about the line.
3. The variances of the extremes (largest and smallest values) are much higher than the variances in the middle of the plot. Greater discrepancies can be accepted at the extremes. The linearity of the points in the middle of the plot is more important than the linearity at the extremes.

Modern data-analysis software often includes tools for generating $q - q$ plots, especially for the normal distribution. The $q - q$ plot can also be used to compare two samples of data to see whether they can be represented by the same distribution (that is, that they are homogeneous). If $x_1, x_2, \ldots, x_n$ are a sample of the random variable X, and $z_1, z_2, \ldots, z_n$ are a sample of the random variable Z, then plotting the ordered values of X versus the ordered values of Z will reveal approximately a straight line if both samples are well represented by the same distribution (Chambers, Cleveland, and Tukey [1983]).

9.3 PARAMETER ESTIMATION

After a family of distributions has been selected, the next step is to estimate the parameters of the distribution. Estimators for many useful distributions are described in this section. In addition, many software packages—some of them integrated into simulation languages—are now available to compute these estimates.

9.3.1 Preliminary Statistics: Sample Mean and Sample Variance

In a number of instances, the sample mean, or the sample mean and sample variance, are used to estimate the parameters of a hypothesized distribution; see Example 9.4. In the following paragraphs, three sets of equations are given for computing the sample mean and sample variance. Equations (9.1) and (9.2) can be used when discrete or continuous raw data are available. Equations (9.3) and (9.4) are used when the data are discrete and have been grouped in a frequency distribution. Equations (9.5) and (9.6) are used when the data are discrete or continuous and have been placed in class intervals. Equations (9.5) and (9.6) are approximations and should be used only when the raw data are unavailable.

If the observations in a sample of size n are $X_1, X_2, \ldots, X_n$, the sample mean ($\bar{X}$) is defined by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{9.1}$$

and the sample variance, S^2, is defined by

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1} \tag{9.2}$$

If the data are discrete and have been grouped in a frequency distribution, Equations (9.1) and (9.2) can be modified to provide for much greater computational efficiency. The sample mean can be computed as

$$\bar{X} = \frac{\sum_{j=1}^{k} f_j X_j}{n} \tag{9.3}$$

and the sample variance as

$$S^2 = \frac{\sum_{j=1}^{k} f_j X_j^2 - n\bar{X}^2}{n-1} \tag{9.4}$$

where k is the number of distinct values of X and f_j is the observed frequency of the value X_j of X.

Example 9.5: Grouped Data

The data in Table 9.1 can be analyzed to obtain $n = 100, f_1 = 12, X_1 = 0, f_2 = 10, X_2 = 1, \ldots, \sum_{j=1}^{k} f_j X_j = 364$, and $\sum_{j=1}^{k} = f_j X_j^2 = 2080$. From Equation (9.3),

$$\bar{X} = \frac{364}{100} = 3.64$$

and, from Equation (9.4),

$$S^2 = \frac{2080 - 100(3.64)^2}{99} = 7.63$$

The sample standard deviation, S, is just the square root of the sample variance. In this case, $S = \sqrt{7.63} = 2.76$. Equations (9.1) and (9.2) would have yielded exactly the same results for $\bar{X}$ and S^2.

It is preferable to use the raw data, if possible, when the values are continuous. However, data sometimes are received after having been placed in class intervals. Then it is no longer possible to obtain the exact sample mean and variance. In such cases, the sample mean and sample variance are approximated from the following equations:

$$\bar{X} \doteq \frac{\sum_{j=1}^{c} f_j m_j}{n} \tag{9.5}$$

and

$$S^2 \doteq \frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n-1} \tag{9.6}$$

where f_j is the observed frequency in the jth class interval, m_j is the midpoint of the jth interval, and c is the number of class intervals.

Example 9.6: Continuous Data in Class Intervals

Assume that the raw data on component life shown in Example 9.3 either was discarded or was lost. However, the data shown in Table 9.2 are still available. To approximate values for $\bar{X}$ and S^2, Equations (9.5) and (9.6) are used. The following values are created: $f_1 = 23$, $m_1 = 1.5$, $f_2 = 10$, $m_2 = 4.5$, …, $\sum_{j=1}^{49} f_j m_j = 614$ and $\sum_{j=1}^{49} f_j m_j^2 = 37{,}226.5$. With $n = 50$, $\bar{X}$ is approximated from Equation (9.5) as

$$\bar{X} \doteq \frac{614}{50} = 12.28$$

Then, S^2 is approximated from Equation (9.6) as

$$S^2 \doteq \frac{37{,}226.5 - 50(12.28)^2}{49} = 605.849$$

and

$$S \doteq 24.614$$

Applying Equations (9.1) and (9.2) to the original data in Example 9.3 results in $\bar{X} = 11.894$ and $S = 24.953$. Thus, when the raw data are either discarded or lost, inaccuracies could result.

9.3.2 Suggested Estimators

Numerical estimates of the distribution parameters are needed to reduce the family of distributions to a specific distribution and to test the resulting hypothesis. Table 9.3 contains suggested estimators for distributions often used in simulation, all of which were described in Chapter 5. Except for an adjustment to remove bias in the estimate of σ^2 for the normal distribution, these estimators are the maximum-likelihood estimators based on the raw data. (If the data are in class intervals, these estimators must be modified.) The reader

Table 9.3 Suggested Estimators for Distributions Often Used in Simulation

Distribution	*Parameter(s)*	*Suggested Estimator(s)*
Poisson	α	$\hat{\alpha} = \bar{X}$
Exponential	λ	$\hat{\lambda} = \dfrac{1}{\bar{X}}$
Gamma	β, θ	$\hat{\beta}$ (see Table A.9) $\hat{\theta} = \dfrac{1}{\bar{X}}$
Normal	μ, σ^2	$\hat{\mu} = \bar{X}$ $\hat{\sigma}^2 = S^2$ (unbiased)
Lognormal	μ, σ^2	$\hat{\mu} = \bar{X}$ (after taking ln of the data) $\hat{\sigma}^2 = S^2$ (after taking ln of the data)
Weibull with $\nu = 0$	α, β	$\hat{\beta}_0 = \dfrac{\bar{X}}{S}$ $\hat{\beta}_j = \hat{\beta}_{j-1} - \dfrac{f(\hat{\beta}_{j-1})}{f'(\hat{\beta}_{j-1})}$ See Equations (9.12) and (9.15) for $f(\hat{\beta})$ and $f'(\hat{\beta})$ Iterate until convergence $\hat{\alpha} = \left(\dfrac{1}{n}\sum_{i=1}^{n} X_i^{\hat{\beta}}\right)^{1/\hat{\beta}}$
Beta	β_1, β_2	$\Psi(\hat{\beta}_1) + \Psi(\hat{\beta}_1 - \hat{\beta}_2) = \ln(G_1)$ $\Psi(\hat{\beta}_2) + \Psi(\hat{\beta}_1 - \hat{\beta}_2) = \ln(G_2)$ where Ψ is the digamma function, $G_1 = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$ and $G_2 = \left(\prod_{i=1}^{n} (1 - X_i)\right)^{1/n}$

is referred to Fishman [1973] and Law and Kelton [2000] for parameter estimates for the uniform, binomial, and negative binomial distributions. The triangular distribution is usually employed when no data are available, with the parameters obtained from educated guesses for the minimum, most likely, and maximum possible values; the uniform distribution may also be used in this way if only minimum and maximum values are available.

Examples of the use of the estimators are given in the following paragraphs. The reader should keep in mind that a parameter is an unknown constant, but the estimator is a statistic (or random variable), because it depends on the sample values. To distinguish the two clearly here, if, say, a parameter is denoted by α, the estimator will be denoted by $\hat{\alpha}$.

Example 9.7: Poisson Distribution

Assume that the arrival data in Table 9.1 require analysis. By comparison with Figure 5.7, an examination of Figure 9.2 suggests a Poisson distributional assumption with unknown parameter α. From Table 9.3, the estimator of α is $\bar{X}$, which was found in Example 9.5. Thus, $\hat{\alpha} = 3.64$. Recall that the true mean and variance are equal for the Poisson distribution. In Example 9.5, the sample variance was estimated as $S^2 = 7.63$. However, it should never be expected that the sample mean and the sample variance will be precisely equal, because each is a random variable.

Example 9.8: Lognormal Distribution

The rates of return on 10 investments in a portfolio are 18.8, 27.9, 21.0, 6.1, 37.4, 5.0, 22.9, 1.0, 3.1 and 8.3 percent. To estimate the parameters of a lognormal model of these data, we first take the natural log of the data and obtain 2.9, 3.3, 3.0, 1.8, 3.6, 1.6, 3.1, 0, 1.1, and 2.1. Then we set $\hat{\mu} = \bar{X} = 2.3$ and $\hat{\sigma}^2 = S^2 = 1.3$·

Example 9.9: Normal Distribution

The parameters of the normal distribution, μ and σ^2, are estimated by $\bar{X}$ and S^2, as shown in Table 9.3. The $q - q$ plot in Example 9.4 leads to a distributional assumption that the installation times are normal. From Equations (9.1) and (9.2), the data in Example 9.4 yield $\hat{\mu} = \bar{X} = 99.9865$ and $\hat{\sigma} = S^2 = (0.2832)^2$ second^2.

Example 9.10: Gamma Distribution

The estimator $\hat{\beta}$ for the gamma distribution is chosen by the use of Table A.9, from Choi and Wette [1969]. Table A.9 requires the computation of the quantity $1/M$, where

$$M = \ln \bar{X} - \frac{1}{n}\sum_{i=1}^{n} \ln X_i \qquad (9.7)$$

Also, it can be seen in Table 9.3 that $\hat{\theta}$ is given by

$$\hat{\theta} = \frac{1}{\bar{X}} \qquad (9.8)$$

In Chapter 5, it was stated that lead time is often gamma distributed. Suppose that the lead times (in days) associated with 20 orders have been accurately measured as follows:

Order	*Lead Time (Days)*	*Order*	*Lead Time (Days)*
1	70.292	11	30.215
2	10.107	12	17.137
3	48.386	13	44.024
4	20.480	14	10.552
5	13.053	15	37.298
6	25.292	16	16.314
7	14.713	17	28.073
8	39.166	18	39.019
9	17.421	19	32.330
10	13.905	20	36.547

To estimate $\hat{\beta}$ and $\hat{\theta}$, it is first necessary to compute M from Equation (9.7). Here, $\overline{X}$ is found, from Equation (9.1), to be

$$\overline{X} = \frac{564.32}{20} = 28.22$$

Then,

$$\ln \overline{X} = 3.34$$

Next,

$$\sum_{i=1}^{20} \ln X_i = 63.99$$

Then,

$$M = 3.34 - \frac{63.99}{20} = 0.14$$

and

$$1/M = 7.14$$

By interpolation in Table A.9, $\hat{\beta} = 3.728$. Finally, Equation (9.8) results in

$$\hat{\theta} = \frac{1}{28.22} = 0.035$$

Example 9.11: Exponential Distribution

Assuming that the data in Example 9.3 come from an exponential distribution, the parameter estimate, $\hat{\lambda}$, can be determined. In Table 9.3, $\hat{\lambda}$ is obtained from $\overline{X}$ as follows:

$$\hat{\lambda} = \frac{1}{\overline{X}} = \frac{1}{11\,894} = 0.084 \text{ per day}$$

Example 9.12: Weibull Distribution

Suppose that a random sample of size n, $X_1, X_2, \ldots, X_n$, has been taken and that the observations are assumed to come from a Weibull distribution. The likelihood function derived by using the pdf given by Equation (5.47) can be shown to be

$$L(\alpha, \beta) = \frac{\beta^n}{\alpha^{\beta n}}\left[\prod_{i=1}^{n} X_i^{(\beta - 1)}\right]\exp\left[-\sum_{i=1}^{n}\left(\frac{X_i}{\alpha}\right)^{\beta}\right] \tag{9.9}$$

The maximum-likelihood estimates are those values of $\hat{\alpha}$ and $\hat{\beta}$ that maximize $L(\alpha, \beta)$ or, equivalently, maximize $\ln L(\alpha, \beta)$, denoted by $l(\alpha, \beta)$. The maximum value of $l(\alpha, \beta)$ is obtained by taking the partial derivatives $\partial l(\alpha, \beta)/\partial\alpha$ and $\partial l(\alpha, \beta)/\partial\beta$, setting each to zero, and solving the resulting equations, which after substitution become

$$f(\beta) = 0 \tag{9.10}$$

and

$$\alpha = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\beta}\right)^{1/\beta} \tag{9.11}$$

where

$$f(\beta) = \frac{n}{\beta} + \sum_{i=1}^{n} \ln X_i - \frac{n\sum_{i=1}^{n} X_i^{\beta} \ln X_i}{\sum_{i=1}^{n} X_i^{\beta}} \tag{9.12}$$

The maximum-likelihood estimates, $\hat{\alpha}$ and $\hat{\beta}$, are the solutions of Equations (9.10) and (9.11). First, $\hat{\beta}$ is found via the iterative procedure explained shortly. Then $\hat{\alpha}$ is found from Equation (9.11), with $\beta = \hat{\beta}$.

Equation (9.10) is nonlinear, so it is necessary to use a numerical-analysis technique to solve it. In Table 9.3, an iterative method for computing $\hat{\beta}$ is given as

$$\hat{\beta}_j = \hat{\beta}_{j-1} - \frac{f(\hat{\beta}_{j-1})}{f'(\hat{\beta}_{j-1})} \tag{9.13}$$

Equation (9.13) employs Newton's method in reaching $\hat{\beta}$, where $\hat{\beta}_j$ is the jth iteration, beginning with an initial estimate for $\hat{\beta}_0$, given in Table 9.3, as follows:

$$\hat{\beta}_0 = \frac{\bar{X}}{S} \tag{9.14}$$

If the initial estimate, $\hat{\beta}_0$, is sufficiently close to the solution $\hat{\beta}$, then $\hat{\beta}_j$ approaches $\hat{\beta}$ as $j \to \infty$. In Newton's method, $\hat{\beta}$ is approached through increments of size $f(\hat{\beta}_{j-1})/f'(\hat{\beta}_{j-1})$. Equation (9.12) is used to compute $f(\hat{\beta}_{j-1})$ and Equation (9.15) is used to compute, $f'(\hat{\beta}_{j-1})$ as follows:

$$f'(\beta) = -\frac{n}{\beta^2} - \frac{n\sum_{i=1}^{n} X_i^{\beta}(\ln X_i)^2}{\sum_{i=1}^{n} X_i^{\beta}} + \frac{n\left(\sum_{i=1}^{n} X_i^{\beta} \ln X_i\right)^2}{\left(\sum_{i=1}^{n} X_i^{\beta}\right)^2} \tag{9.15}$$

Equation (9.15) can be derived from Equation (9.12) by differentiating $f(\beta)$ with respect to β. The iterative process continues until $f(\hat{\beta}_j) \doteq 0$, for example, until $\left|f(\hat{\beta}_j)\right| \le 0.001$.

Consider the data given in Example 9.3. These data concern the failure of electronic components and looks to come from an exponential distribution. In Example 9.11, the parameter $\hat{\lambda}$ was estimated on the hypothesis that the data were from an exponential distribution. If the hypothesis that the data came from an exponential distribution is rejected, an alternative hypothesis is that the data come from a Weibull distribution. The Weibull distribution is suspected because the data pertain to electronic component failures, which occur suddenly.

Equation (9.14) is used to compute $\hat{\beta}_0$. For the data in Example 9.3, $n = 50$, $\bar{X} = 11.894$, $\bar{X}^2 = 141.467$, and $\sum_{i=1}^{50} X_i^2 = 37{,}575.850$; so S^2 is found by Equation (9.2) to be

$$S^2 = \frac{37{,}578.850 - 50(141.467)}{49} = 622.650$$

and $S = 24.953$. Thus,

$$\hat{\beta}_0 = \frac{11.894}{24.953} = 0.477$$

To compute $\hat{\beta}_1$ by using Equation (9.13) requires the calculation of $f(\hat{\beta}_0)$ and $f'(\hat{\beta}_0)$ from Equations (9.12) and (9.15). The following additional values are needed: $\sum_{i=1}^{50} X_i^{\hat{\beta}_0} = 115.125$, $\sum_{i=1}^{50} \ln X_i = 38.294$, $\sum_{i=1}^{50} X_i^{\hat{\beta}_0} \ln X_i = 292.629$, and $\sum_{i=1}^{50} X_i^{\hat{\beta}_0} (\ln X_i)^2 = 1057.781$. Thus,

$$f(\hat{\beta}_0) = \frac{50}{0.477} + 38.294 - \frac{50(292.629)}{115.125} = 16.024$$

and

$$f'(\hat{\beta}_0) = \frac{-50}{(0.477)^2} - \frac{50(1057.781)}{115.125} + \frac{50(292.629)^2}{(115.125)^2} = -356.110$$

Then, by Equation (9.13),

$$\hat{\beta}_1 = 0.477 - \frac{16.024}{-356.110} = 0.522$$

After four iterations, $|f(\hat{\beta}_3)| \le 0.001$, at which point $\hat{\beta} \doteq \hat{\beta}_4 = 0.525$ is the approximate solution to Equation (9.10). Table 9.4 contains the values needed to complete each iteration.

Now, $\hat{\alpha}$ can be computed from Equation (9.11) with $\beta = \hat{\beta} = 0.525$, as follows:

$$\hat{\alpha} = \left[\frac{130.608}{50}\right]^{1/0.525} = 6.227$$

If $\hat{\beta}_0$ is sufficiently close to $\hat{\beta}$, the procedure converges quickly, usually in four to five iterations. However, if the procedure appears to be diverging, try other initial guesses for $\hat{\beta}_0$—for example, one-half the initial estimate or twice the initial estimate.

The difficult task of estimating parameters for the Weibull distribution by hand emphasizes the value of having software support for input modeling.

Table 9.4 Iterative Estimation of Parameters of the Weibull Distribution

j	$\hat{\beta}_j$	$\sum_{i=1}^{50} X_i^{\hat{\beta}_j}$	$\sum_{i=1}^{50} X_i^{\hat{\beta}_j} \ln X_i$	$\sum_{i=1}^{50} X_i^{\hat{\beta}_j} (\ln X_i)^2$	$f(\hat{\beta}_j)$	$f'(\hat{\beta}_j)$	$\hat{\beta}_{j+1}$
0	0.477	115.125	292.629	1057.781	16.024	−356.110	0.522
1	0.522	129.489	344.713	1254.111	1.008	−313.540	0.525
2	0.525	130.603	348.769	1269.547	0.004	−310.853	0.525
3	0.525	130.608	348.786	1269.614	0.000	−310.841	0.525

```
betaMLE := proc(X, n)
             local G1, G2, beta1, beta2, eqns, solns;
             G1 := product(X[i], i=1..n)^(1/n);
             G2 := product(1-X[i],i=1..n)^(1/n);
             eqns := {Psi(beta1) - Psi(beta1 + beta2) = ln(G1),
                      Psi(beta2) - Psi(beta1 + beta2) = ln(G2)};
             solns := fsolve(eqns, {beta1=0..infinity, beta2=0..infinity});
             RETURN(solns);
             end;
```

Figure 9.5 Maple procedure to compute the maximum likelihood estimates for the beta distribution parameters.

Example 9.13: Beta Distribution

The percentage of customers each month who bring in store coupons must be between 0 and 100 percent. Observations at a store for eight months gave the values 25%, 74%, 20%, 32%, 81%, 47%, 31%, and 8%. To fit a beta distribution to these data, we first need to rescale it to the interval (0, 1) by dividing all the values by 100, to get 0.25, 0.74, 0.20, 0.32, 0.81, 0.47, 0.31, 0.08.

The maximum-likelihood estimators of the parameters β_1, β_2 solve the system of equations shown in Table 9.3. Such equations can be solved by modern symbolic/numerical calculation programs, such as Maple; a Maple procedure for the beta parameters is shown in Figure 9.5. In this case, the solutions are $\hat{\beta}_1 = 1.47$ and $\hat{\beta}_2 = 2.16$.

9.4 GOODNESS-OF-FIT TESTS

Hypothesis testing was discussed in Section 7.4 with respect to testing random numbers. In Section 7.4.1, the Kolmogorov–Smirnov test and the chi-square test were introduced. These two tests are applied in this section to hypotheses about distributional forms of input data.

Goodness-of-fit tests provide helpful guidance for evaluating the suitability of a potential input model; however, there is no single correct distribution in a real application, so you should not be a slave to the verdict of such a test. It is especially important to understand the effect of sample size. If very little data are available, then a goodness-of-fit test is unlikely to reject *any* candidate distribution; but if a lot of data are available, then a goodness-of-fit test will likely reject *all* candidate distributions. Therefore, failing to reject a candidate distribution should be taken as one piece of evidence in favor of that choice, and rejecting an input model as only one piece of evidence against the choice.

9.4.1 Chi-Square Test

One procedure for testing the hypothesis that a random sample of size n of the random variable X follows a specific distributional form is the chi-square goodness-of-fit test. This test formalizes the intuitive idea of comparing the histogram of the data to the shape of the candidate density or mass function. The test is valid for large sample sizes and for both discrete and continuous distributional assumptions when parameters are estimated by maximum likelihood. The test procedure begins by arranging the n observations into a set of k class intervals or cells. The test statistic is given by

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{9.16}$$

where O_i is the observed frequency in the ith class interval and E_i is the expected frequency in that class interval. The expected frequency for each class interval is computed as $E_i = np_i$, where p_i is the theoretical, hypothesized probability associated with the ith class interval.

It can be shown that χ_0^2 approximately follows the chi-square distribution with $k - s - 1$ degrees of freedom, where s represents the number of parameters of the hypothesized distribution estimated by the sample statistics. The hypotheses are the following:

H_0: The random variable, X, conforms to the distributional assumption with the parameter(s) given by the parameter estimate(s).

H_1: The random variable X does not conform.

The critical value $\chi^2_{\alpha,k-s-1}$ is found in Table A.6. The null hypothesis, H_0, is rejected if $\chi_0^2 > \chi^2_{\alpha,k-s-1}$.

When applying the test, if expected frequencies are too small, χ_0^2 will reflect not only the departure of the observed from the expected frequency, but also the smallness of the expected frequency as well. Although there is no general agreement regarding the minimum size of E_i, values of 3, 4, and 5 have been widely used. In Section 7.4.1, when the chi-square test was discussed, the minimum expected frequency five was suggested. If an E_i value is too small, it can be combined with expected frequencies in adjacent class intervals. The corresponding O_i values should also be combined, and k should be reduced by one for each cell that is combined.

If the distribution being tested is discrete, each value of the random variable should be a class interval, unless it is necessary to combine adjacent class intervals to meet the minimum-expected-cell-frequency requirement. For the discrete case, if combining adjacent cells is not required,

$$p_i = p(x_i) = P(X = x_i)$$

Otherwise, p_i is found by summing the probabilities of appropriate adjacent cells.

If the distribution being tested is continuous, the class intervals are given by $[a_{i-1}, a_i)$, where a_{i-1} and a_i are the endpoints of the ith class interval. For the continuous case with assumed pdf $f(x)$, or assumed cdf $F(x)$, p_i can be computed as

$$p_i = \int_{a_{i-1}}^{a_i} f(x)dx = F(a_i) - F(a_{i-1})$$

For the discrete case, the number of class intervals is determined by the number of cells resulting after combining adjacent cells as necessary. However, for the continuous case, the number of class intervals must be specified. Although there are no general rules to be followed, the recommendations in Table 9.5 are made to aid in determining the number of class intervals for continuous data.

Table 9.5 Recommendations for Number of Class Intervals for Continuous Data

Sample Size, n	*Number of Class Intervals,* k
20	Do not use the chi-square test
50	5 to 10
100	10 to 20
>100	$\sqrt{n}$ to $n/5$

Example 9.14: Chi-Square Test Applied to Poisson Assumption

In Example 9.7, the vehicle-arrival data presented in Example 9.2 were analyzed. The histogram of the data, shown in Figure 9.2, appeared to follow a Poisson distribution; hence the parameter, $\hat{\alpha} = 3.64$, was found. Thus, the following hypotheses are formed:

H_0: the random variable is Poisson distributed.
H_1: the random variable is not Poisson distributed.

The pmf for the Poisson distribution was given in Equation (5.19):

$$p(x) = \begin{cases} \dfrac{e^{-\alpha}\alpha^x}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases} \tag{9.17}$$

For $\alpha = 3.64$, the probabilities associated with various values of x are obtained from Equation (9.17):

p(0) = 0.026	p(6) = 0.085
p(1) = 0.096	p(7) = 0.044
p(2) = 0.174	p(8) = 0.020
p(3) = 0.211	p(9) = 0.008
p(4) = 0.192	p(10) = 0.003
p(5) = 0.140	p(≥11) = 0.001

From this information, Table 9.6 is constructed. The value of E_1 is given by $np_0 = 100(0.026) = 2.6$. In a similar manner, the remaining E_i values are computed. Since $E_1 = 2.6 < 5$, E_1 and E_2 are combined. In that case, O_1 and O_2 are also combined, and k is reduced by one. The last five class intervals are also combined, for the same reason, and k is further reduced by four.

The calculated χ_0^2 is 27.68. The degrees of freedom for the tabulated value of χ^2 is $k - s - 1 = 7 - 1 - 1 = 5$. Here, $s = 1$, since one parameter, $\hat{\alpha}$ was estimated from the data. At the 0.05 level of significance, the critical value $\chi^2_{0.05,5}$ is 11.1. Thus, H_0 would be rejected at level of significance 0.05. The analyst, therefore, might want to search for a better-fitting model or use the empirical distribution of the data.

Table 9.6 Chi-square Goodness-of-Fit Test for Example 9.14

x_i	*Observed Frequency,* O_i	*Expected Frequency,* E_i	$\dfrac{(O_i - E_i)^2}{E_i}$
0	12 } 22	2.6 } 12.2	7.87
1	10	9.6	
2	19	17.4	0.15
3	17	21.1	0.80
4	10	19.2	4.41
5	8	14.0	2.57
6	7	8.5	0.26
7	5 } 17	4.4 } 7.6	11.62
8	5	2.0	
9	3	0.8	
10	3	0.3	
≥ 11	1	0.1	
	100	100.0	27.68

9.4.2 Chi-Square Test with Equal Probabilities

If a continuous distributional assumption is being tested, class intervals that are equal in probability rather than equal in width of interval should be used. This has been recommended by a number of authors [Mann and Wald, 1942; Gumbel, 1943; Law and Kelton, 2000; Stuart, Ord, and Arnold, 1998]. It should be noted that the procedure is not applicable to data collected in class intervals, where the raw data have been discarded or lost.

Unfortunately, there is as yet no method for figuring out the probability associated with each interval that maximizes the power for a test of a given size. (The power of a test is defined as the probability of rejecting a false hypothesis.) However, if using equal probabilities, then $pi = 1/k$. We recommend

$$E_i = np_i \geq 5$$

so substituting for p_i yields

$$\frac{n}{k} \geq 5$$

and solving for k yields

$$k \leq \frac{n}{5} \tag{9.18}$$

Equation (9.18) was used in coming up with the recommendations for maximum number of class intervals in Table 9.5.

If the assumed distribution is normal, exponential, or Weibull, the method described in this section is straightforward. Example 9.15 indicates how the procedure is accomplished for the exponential distribution. If the assumed distribution is gamma (but not Erlang) or certain other distributions, then the computation of endpoints for class intervals is complex and could require numerical integration of the density function. Statistical-analysis software is very helpful in such cases.

Example 9.15: Chi-Square Test for Exponential Distribution

In Example 9.11, the failure data presented in Example 9.3 were analyzed. The histogram of the data, shown in Figure 9.3, appeared to follow an exponential distribution, so the parameter $\hat{\lambda} = 1/\bar{X} = 0.084$ was computed. Thus, the following hypotheses are formed:

H_0: the random variable is exponentially distributed.
H_1: the random variable is not exponentially distributed.

In order to perform the chi-square test with intervals of equal probability, the endpoints of the class intervals must be found. Equation (9.18) indicates that the number of intervals should be less than or equal to $n/5$. Here, $n = 50$, and so $k \leq 10$. In Table 9.5, it is recommended that 7 to 10 class intervals be used. Let $k = 8$; then each interval will have probability $p = 0.125$. The endpoints for each interval are computed from the cdf for the exponential distribution, given in Equation (5.28), as follows:

$$F(a_i) = 1 - e^{-\lambda a_i} \tag{9.19}$$

where a_i represents the endpoint of the ith interval, $i = 1, 2, \ldots, k$. Since $F(a_i)$ is the cumulative area from zero to a_i, $F(a_i) = ip$, so Equation (9.19) can be written as

$$ip = 1 - e^{-\lambda a_i}$$

or

$$e^{-\lambda a_i} = 1 - ip$$

Taking the logarithm of both sides and solving for a_i gives a general result for the endpoints of k equiprobable intervals for the exponential distribution:

$$a_i = -\frac{1}{\lambda}\ln(1 - ip), \qquad i = 0, 1, \ldots, k \tag{9.20}$$

Regardless of the value of λ, Equation (9.20) will always result in $a_0 = 0$ and $a_k = \infty$. With $\hat{\lambda} = 0.084$ and $k = 8$, a_1 is computed from Equation (9.20) as

$$a_1 = -\frac{1}{0.084}\ln(1 - 0.125) = 1.590$$

Continued application of Equation (9.20) for $i = 2, 3, \ldots, 7$ results in $a_2, \ldots, a_7$ as 3.425, 5.595, 8.252, 11.677, 16.503, and 24.755. Since $k = 8$, $a_8 = \infty$. The first interval is [0, 1.590), the second interval is [1.590, 3.425), and so on. The expectation is that 0.125 of the observations will fall in each interval. The observations, the expectations, and the contributions to the calculated value of χ_0^2 are shown in Table 9.7. The calculated value of χ_0^2 is 39.6. The degrees of freedom are given by $k - s - 1 = 8 - 1 - 1 = 6$. At $\alpha = 0.05$, the tabulated value of $\chi^2_{0.05,6}$ is 12.6. Since $\chi_0^2 > \chi^2_{0.05,6}$, the null hypothesis is rejected. (The value of $\chi^2_{0.01,6}$ is 16.8, so the null hypothesis would also be rejected at level of significance $\alpha = 0.01$.)

Table 9.7 Chi-Square Goodness-of-Fit Test for Example 9.15

Class Interval	*Observed Frequency,* O_i	*Expected Frequency,* E_i	$\frac{(O_i - E_i)^2}{E_i}$
[0, 1.590)	19	6.25	26.01
[1.590, 3.425)	10	6.25	2.25
[3.425, 5.595)	3	6.25	0.81
[5.595, 8.252)	6	6.25	0.01
[8.252, 11.677)	1	6.25	4.41
[11.677, 16.503)	1	6.25	4.41
[16.503, 24.755)	4	6.25	0.81
[24.755, ∞)	6	6.25	0.01
	50	50	39.6

9.4.3 Kolmogorov–Smirnov Goodness-of-Fit Test

The chi-square goodness-of-fit test can accommodate the estimation of parameters from the data with a resultant decrease in the degrees of freedom (one for each parameter estimated). The chi-square test requires that the data be placed in class intervals; in the case of a continuous distributional assumption, this grouping is arbitrary. Changing the number of classes and the interval width affects the value of the calculated and tabulated chi-square. A hypothesis could be accepted when the data are grouped one way, but rejected when they are grouped another way. Also, the distribution of the chi-square test statistic is known only approximately, and the power of the test is sometimes rather low. As a result of these considerations, goodness-of-fit tests other than the chi-square, are desired. The Kolmogorov–Smirnov test formalizes the idea behind examining a $q - q$ plot.

The Kolmogorov–Smirnov test was presented in Section 7.4.1 to test for the uniformity of numbers. Both of these uses fall into the category of testing for goodness of fit. Any continuous distributional assumption can be tested for goodness of fit by using the method of Section 7.4.1.

The Kolmogorov–Smirnov test is particularly useful when sample sizes are small and when no parameters have been estimated from the data. When parameter estimates have been made, the critical values in Table A.8 are biased; in particular, they are too conservative. In this context, "conservative" means that the critical values will be too large, resulting in smaller Type I (α) errors than those specified. The exact value of α can be worked out in some instances, as is discussed at the end of this section.

The Kolmogorov–Smirnov test does not take any special tables when an exponential distribution is assumed. The following example indicates how the test is applied in this instance. (Notice that it is not necessary to estimate the parameter of the distribution in this example, so we may use Table A.8.)

Example 9.16: Kolmogorov–Smirnov Test for Exponential Distribution

Suppose that 50 interarrival times (in minutes) are collected over the following 100-minute interval (arranged in order of occurrence):

0.44	0.53	2.04	2.74	2.00	0.30	2.54	0.52	2.02	1.89	1.53	0.21
2.80	0.04	1.35	8.32	2.34	1.95	0.10	1.42	0.46	0.07	1.09	0.76
5.55	3.93	1.07	2.26	2.88	0.67	1.12	0.26	4.57	5.37	0.12	3.19
1.63	1.46	1.08	2.06	0.85	0.83	2.44	1.02	2.24	2.11	3.15	2.90
6.58	0.64										

The null hypothesis and its alternate are formed as follows:

H_0: the interarrival times are exponentially distributed.
H_1: the interarrival times are not exponentially distributed.

The data were collected over the interval from 0 to $T = 100$ minutes. It can be shown that, if the underlying distribution of interarrival times $\{T_1, T_2, \ldots\}$ is exponential, the arrival times are uniformly distributed on the interval $(0, T)$. The arrival times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, ..., $T_1 + \cdots + T_{50}$ are obtained by adding interarrival times. The arrival times are then normalized to a (0, 1) interval so that the Kolmogorov–Smirnov test, as presented in Section 7.4.1, can be applied. On a (0, 1) interval, the points will be [T_1/T, $(T_1 + T_2)/T$, ..., $(T_1 + \cdots + T_{50})/T$]. The resulting 50 data points are as follows:

0.0044	0.0097	0.0301	0.0575	0.0775	0.0805	0.1059	0.1111	0.1313	0.1502
0.1655	0.1676	0.1956	0.1960	0.2095	0.2927	0.3161	0.3356	0.3366	0.3508
0.3553	0.3561	0.3670	0.3746	0.4300	0.4694	0.4796	0.5027	0.5315	0.5382
0.5494	0.5520	0.5977	0.6514	0.6526	0.6845	0.7008	0.7154	0.7262	0.7468
0.7553	0.7636	0.7880	0.7982	0.8206	0.8417	0.8732	0.9022	0.9680	0.9744

Following the procedure in Example 7.6 produces a D^+ of 0.1054 and a D^- of 0.0080. Therefore, the Kolmogorov–Smirnov statistic is $D = \max(0.1054, 0.0080) = 0.1054$. The critical value of D obtained from Table A.8 for a level of significance of $\alpha = 0.05$ and $n = 50$ is $D_{0.05} = 1.36/\sqrt{n} = 0.1923$; but $D = 0.1054$, so the hypothesis that the interarrival times are exponentially distributed cannot be rejected.

The Kolmogorov–Smirnov test has been modified so that it can be used in several situations where the parameters are estimated from the data. The computation of the test statistic is the same, but different tables of critical values are used. Different tables of critical values are required for different distributional assumptions. Lilliefors [1967] developed a test for normality. The null hypothesis states that the population is one of the family of normal distributions, without specifying the parameters of the distribution. The interested reader might wish to study Lilliefors' original work; he describes how simulation was used to develop the critical values.

Lilliefors [1969] also modified the critical values of the Kolmogorov–Smirnov test for the exponential distribution. Lilliefors again used random sampling to obtain approximate critical values, but Durbin [1975] subsequently obtained the exact distribution. Connover [1998] gives examples of Kolmogorov–Smirnov tests for the normal and exponential distributions. He also refers to several other Kolmogorov–Smirnov-type tests that might be of interest to the reader.

A test that is similar in spirit to the Kolmogorov–Smirnov test is the *Anderson–Darling test.* Like the Kolmogorov–Smirnov test, the Anderson–Darling test is based on the difference between the empirical cdf and the fitted cdf; unlike the Kolmogorov–Smirnov test, the Anderson–Darling test is based on a more comprehensive measure of difference (not just the maximum difference) and is more sensitive to discrepancies in the tails of the distributions. The critical values for the Anderson–Darling test also depend on the candidate distribution and on whether parameters have been estimated. Fortunately, this test and the Kolmogorov–Smirnov test have been implemented in a number of software packages that support simulation-input modeling.

9.4.4 *p*-Values and "Best Fits"

To apply a goodness-of-fit test, a significance level must be chosen. Recall that the significance level is the probability of falsely rejecting H_0: the random variable conforms to the distributional assumption. The traditional significance levels are 0.1, 0.05 and 0.01. Prior to the availability of high-speed computing, having a small set of standard values made it possible to produce tables of useful critical values. Now most statistical software computes critical values as needed, rather than storing them in tables. Thus, the analyst can employ a different level of significance—say, 0.07.

However, rather than require a prespecified significance level, many software packages compute a *p-value* for the test statistic. The p-value is the significance level at which one would *just reject* H_0 for the given value of the test statistic. Therefore, a large p-value tends to indicate a good fit (we would have to accept a large chance of error in order to reject), while a small p-value suggests a poor fit (to accept we would have to insist on almost no risk).

Recall Example 9.14, in which a chi-square test was used to check the Poisson assumption for the vehicle-arrival data. The value of the test statistic was $\chi_0^2 = 27.58$, with 5 degrees of freedom. The p-value for this test statistic is 0.00004, meaning that we would reject the hypothesis that the data are Poisson at the 0.00004 significance level. (Recall that we rejected the hypothesis at the 0.05 level; now we know that we would also to reject it at even lower levels.)

The p-value can be viewed as a measure of fit, with larger values being better. This suggests that we could fit every distribution at our disposal, compute a test statistic for each fit, and then choose the distribution that yields the largest p-value. We know of no input modeling software that implements this specific algorithm, but many such packages do include a "best fit" option, in which the software recommends an input model to the user after evaluating all feasible models. The software might also take into account other factors—such as whether the data are discrete or continuous, bounded or unbounded—but, in the end, some summary measure of fit, like the p-value, is used to rank the distributions. There is nothing wrong with this, but there are several things to keep in mind:

1. The software might know nothing about the physical basis of the data, whereas that information can suggest distribution families that are appropriate. (See the list in Section 9.2.2.) Remember that the goal of input modeling is often to fill in gaps or smooth the data, rather than find an input model that conforms as closely as possible to the given sample.
2. Recall that both the Erlang and the exponential distributions are special cases of the gamma and that the exponential is also a special case of the more flexible Weibull. Automated best-fit procedures tend to choose the more flexible distributions (gamma and Weibull over Erlang and exponential), because the extra flexibility allows closer conformance to the data and a better summary measure of fit. But again, close conformance to the data does not always lead to the most appropriate input model.
3. A summary statistic, like the p-value, is just that, a summary measure. It says little or nothing about where the lack of fit occurs (in the body of the distribution, in the right tail, or in the left tail). A human, using graphical tools, can see where the lack of fit occurs and decide whether or not it is important for the application at hand.

Our recommendation is that automated distribution selection be used as one of several ways to suggest candidate distributions. Always inspect the automatic selection, using graphical methods, and remember that the final choice is yours.

9.5 FITTING A NONSTATIONARY POISSON PROCESS

Fitting a nonstationary Poisson process (NSPP) to arrival data is a difficult problem, in general, because we seldom have knowledge about the appropriate form of the arrival rate function $\lambda(t)$. (See Chapter 5, Section 5.5 for the definition of a NSPP). One approach is to choose a very flexible model with lots of parameters and fit it with a method such as maximum likelihood: see Johnson, Lee, and Wilson [1994] for an example of this approach. A second method, and the one we consider here, is to approximate the arrival rate as being constant over some basic interval of time, such as an hour, or a day, or a month, but varying from time interval to time interval. The problem then becomes choosing the basic time interval and estimating the arrival rate within each interval.

Suppose we need to model arrivals over a time period, say $[0, T]$. The approach that we describe is most appropriate when it is possible to observe the time period $[0, T]$ repeatedly and count arrivals. For instance, if the problem involves modeling the arrival of e-mail throughout the business day (8 A.M. to 6 P.M.), and we believe that the arrival rate is approximately constant over half-hour intervals, then we need to be able to count arrivals during half-hour intervals for several days. If it is possible to record actual arrival times, rather than counts, then actual arrival times are clearly better since they can later be grouped into any interval lengths we desire. However, we will assume from here on that only counts are available.

Divide the time period $[0, T]$ into k equal intervals of length $\Delta t = T/k$. For instance, if we are considering a 10-hour business day from 8 A.M. to 6 P.M. and if we allow the rate to change every half hour, then $T = 10$, $k = 20$, and $\Delta t = 1/2$. Over n periods of observation (e.g., n days), let C_{ij} be the number of arrivals that occurred during the ith time interval on the jth period of observation. In our example, C_{23} would be the number of arrivals from 8:30 A.M. to 9 A.M. (second half-hour period) on the third day of observation.

The estimated arrival rate during the ith time period, $(i-1)\Delta t < t \leq i\,\Delta t$, is then just the average number of arrivals scaled by the length of the time interval:

$$\hat{\lambda}(t) = \frac{1}{n\Delta t}\sum_{j=1}^{n} C_{ij} \tag{9.21}$$

Table 9.8 Monday E-mail Arrival Data for NSPP Example

Time Period	*Number of Arrivals*			*Estimated Arrival Rate (arrivals/hour)*
	Day 1	*Day 2*	*Day 3*	
8:00–8:30	12	14	10	24
8:30–9:00	23	26	32	54
9:00–9:30	27	19	32	52
9:30–10:00	20	13	12	30

After the arrival rates for each time interval have been estimated, adjacent intervals whose rates appear to be the same can be combined.

For instance, consider the e-mail arrival counts during the first two hours of the business day on three Mondays, shown in Table 9.8. The estimated arrival rate for 8:30–9:00 is

$$\frac{1}{3(1/2)}(23+26+32) = 54 \text{ arrivals/hour}$$

After seeing these results we might consider combining the interval 8:30–9:00 with the interval 9:00–9:30, because the rates are so similar. Note also that the goodness-of-fit tests described in the previous section can be applied to the data from each time interval individually, to check the Poisson approximation.

9.6 SELECTING INPUT MODELS WITHOUT DATA

Unfortunately, it is often necessary in practice to develop a simulation model—perhaps for demonstration purposes or a preliminary study—before any process data are available. In this case, the modeler must be resourceful in choosing input models and must carefully check the sensitivity of results to the chosen models.

There are a number of ways to obtain information about a process even if data are not available:

Engineering data: Often a product or process has performance ratings provided by the manufacturer (for example, the mean time to failure of a disk drive is 10000 hours; a laser printer can produce 8 pages/minute; the cutting speed of a tool is 1 cm/second; etc.). Company rules might specify time or production standards. These values provide a starting point for input modeling by fixing a central value.

Expert option: Talk to people who are experienced with the process or similar processes. Often, they can provide optimistic, pessimistic, and most-likely times. They might also be able to say whether the process is nearly constant or highly variable, and they might be able to define the source of variability.

Physical or conventional limitations: Most real processes have physical limits on performance—for example, computer data entry cannot be faster than a person can type. Because of company policies, there could be upper limits on how long a process may take. Do not ignore obvious limits or bounds that narrow the range of the input process.

The nature of the process: The description of the distributions in Section 9.2.2 can be used to justify a particular choice even when no data are available.

When data are not available, the uniform, triangular, and beta distributions are often used as input models. The uniform can be a poor choice, because the upper and lower bounds are rarely just as likely as the central values in real processes. If, in addition to upper and lower bounds, a most-likely value can be given, then the triangular distribution can be used. The triangular distribution places much of its probability near the most-likely value, and much less near the extremes. (See Section 5.4.) If a beta distribution is used, then be sure to plot the density function of the selected distribution; the beta can take unusual shapes.

A useful refinement is obtained when a minimum, a maximum, and one or more "breakpoints" can be given. A breakpoint is an intermediate value together with a probability of being less than or equal to that value. The following example illustrates how breakpoints are used.

Example 9.17

For a production-planning simulation, the sales volume of various products is required. The salesperson responsible for product XYZ-123 says that no fewer than 1000 units will be sold (because of existing contracts) and no more than 5000 units will be sold (because that is the entire market for the product). Given her experience, she believes that there is a 90% chance of selling more than 2000 units, a 25% chance of selling more than 3500 units, and only a 1% chance of selling more than 4500 units.

Table 9.9 summarizes this information. Notice that the chances of exceeding certain sales goals have been translated into the cumulative probability of being less than or equal to those goals. With the information in this form, the method of Section 8.1.5 can be employed to generate simulation-input data.

When input models have been selected without data, it is especially important to test the sensitivity of simulation results to the distribution chosen. Check sensitivity not only to the center of the distribution, but also to the variability or limits. Extreme sensitivity of output results to the input model provides a convincing argument against making critical decisions based on the results and in favor of undertaking data collection.

For additional discussion of input modeling in the absence of data, see Pegden, Shannon, and Sadowski [1995].

9.7 MULTIVARIATE AND TIME-SERIES INPUT MODELS

In Sections 9.1–9.4, the random variables presented were considered to be independent of any other variables within the context of the problem. However, variables may be related, and, if the variables appear in a simulation model as inputs, the relationship should be investigated and taken into consideration.

Example 9.18

An inventory simulation includes the lead time and annual demand for industrial robots. An increase in demand results in an increase in lead time: The final assembly of the robots must be made according to the specifications of the purchaser. Therefore, rather than treat lead time and demand as independent random variables, a multivariate input model should be developed.

Table 9.9 Summary of Sales Information

i	*Interval (Sales)*	*Cumulative Frequency, c_i*
1	$1000 \le x \le 2000$	0.10
2	$2000 < x \le 3500$	0.75
3	$3500 < x \le 4500$	0.99
4	$4500 < x \le 5000$	1.00