22.7 VARIANCE-REDUCING TECHNIQUES

Because considerable computer time usually is required for simulation runs, it is important to obtain as much and as precise information as possible from the amount of simulation that can be done. Unfortunately, there has been a tendency in practice to apply simulation uncritically without giving adequate thought to the efficiency of the experimental design. This tendency has occurred despite the fact that considerable progress has been made in developing special techniques for increasing the precision (i.e., decreasing the variance) of sample estimators.

These variance-reducing techniques often are called Monte Carlo techniques (a term sometimes applied to simulation in general). Because they tend to be rather sophisticated, it is not possible to explore them deeply here. However, we shall attempt to impart the flavor of these techniques and the great increase in precision they sometimes provide by presenting two when applied to the following example.

Consider the exponential distribution whose parameter has a value of 1. Thus, its probability density function is $f(x)\quad e^{x}$, as shown in Fig. 22.15, and its cumulative distribution function is $F(x)\quad 1\quad e^{x}$. It is known that the mean of this distribution is 1. However, suppose that this mean is not known and that we want to estimate this mean by using simulation.

To provide a standard of comparison of the two variance-reducing techniques, we consider first the straightforward simulation approach, sometimes called the crude Monte Carlo technique. This approach involves generating some *random observations* from the

FIGURE 22.15 Probability density function for the example for variance-reducing techniques, where the objective is to estimate the mean of this distribution.

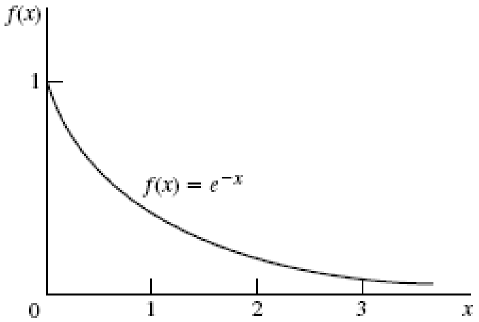
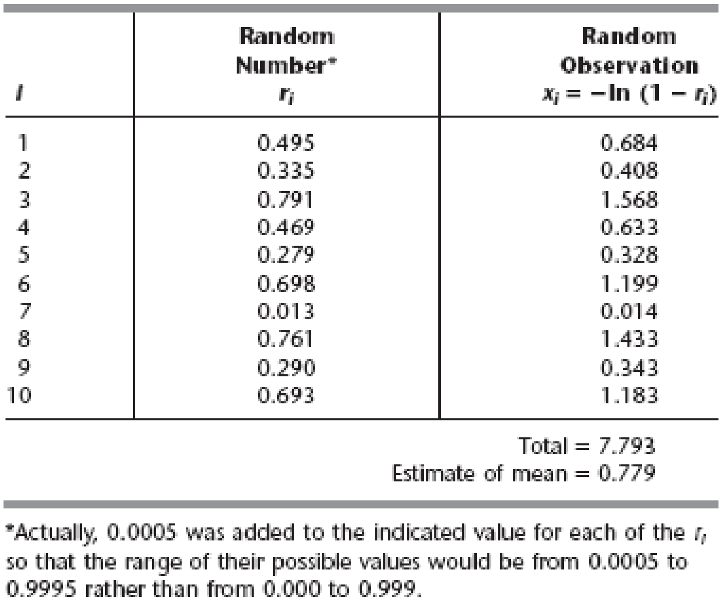


TABLE 22.7 Application of the crude Monte Carlo technique to the example



i	Random Number* r_i	Random Observation $x_i = -\ln(1 - r_i)$
1	0.495	0.684
2	0.335	0.408
3	0.791	1.568
4	0.469	0.633
5	0.279	0.328
6	0.698	1.199
7	0.013	0.014
8	0.761	1.433
9	0.290	0.343
10	0.693	1.183
		Total = 7.793
		Estimate of mean = 0.779

*Actually, 0.0005 was added to the indicated value for each of the r_i so that the range of their possible values would be from 0.0005 to 0.9995 rather than from 0.000 to 0.999.

exponential distribution under consideration and then using the *average* of these observations to estimate the mean. As described in Sec. 22.4, these random observations would be

xi ln (1 *ri*), for *i* 1, 2, . . . , *n*,

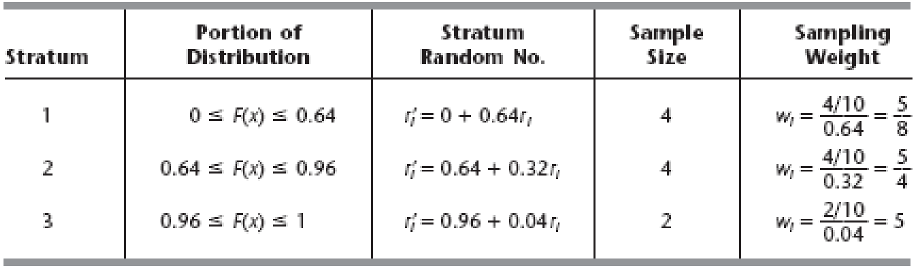
where *r*1, *r*2, . . . , *rn* are uniform random numbers between 0 and 1. We use the first three digits in the fifth column of Table 22.3 to obtain 10 such uniform random numbers; the resulting random observations are shown in Table 22.7. (These same random numbers also are used to illustrate the variance-reducing techniques to sharpen the comparison.)

Notice that the sample average in Table 22.7 is 0.779, as opposed to the true mean of 1.000. However, because the standard deviation of the sample average happens to be 1/*n*, or 1/10 in this case (as could be estimated from the sample), an error of this amount or larger would occur approximately one-half of the time. Furthermore, because the standard deviation of a sample average is always inversely proportional to *n*, this sample size would need to be quadrupled to reduce this standard deviation by one-half. These somewhat disheartening facts suggest the need for other techniques that would obtain such estimates more precisely and more efficiently.

Stratified Sampling

Stratified sampling is a relatively simple Monte Carlo technique for obtaining better estimates. There are two shortcomings of the crude Monte Carlo approach that are rectified by stratified sampling. First, by the very nature of randomness, a random sample may not provide a particularly uniform cross section of the distribution. For example, the random sample given in Table 22.7 has no observations between 0.014 and 0.328, even though the probability that a random observation will fall inside this interval is greater than .

TABLE 22.8 Formulation of the stratified sampling approach to the example



Stratum	Portion of Distribution	Stratum Random No.	Sample Size	Sampling Weight
1	$0 \leq F(x) \leq 0.64$	$r_i' = 0 + 0.64r_i$	4	$w_i = \frac{4/10}{0.64} = \frac{5}{8}$
2	$0.64 \leq F(x) \leq 0.96$	$r_i' = 0.64 + 0.32r_i$	4	$w_i = \frac{4/10}{0.32} = \frac{5}{4}$
3	$0.96 \leq F(x) \leq 1$	$r_i' = 0.96 + 0.04r_i$	2	$w_i = \frac{2/10}{0.04} = 5$

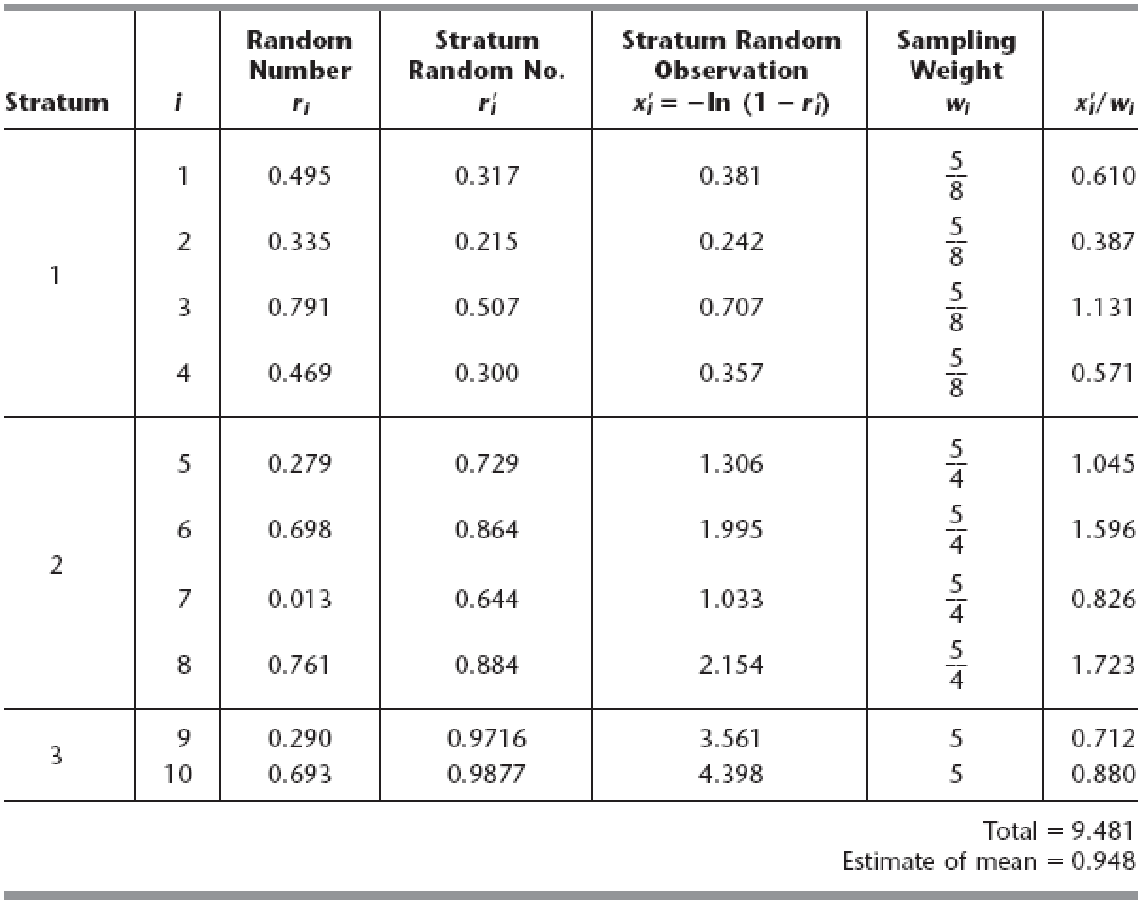
Second, certain portions of a distribution may be more critical than others for obtaining a precise estimate, but random sampling gives no special priority to obtaining observations from these portions. For example, the tail of an exponential distribution is especially critical in determining its mean. However, the random sample in Table 22.7 includes no observations larger than 1.568, even though there is at least a small probability of *much* larger values. This explanation is the basic one for why this particular sample average is far below the true mean. Stratified sampling circumvents these difficulties by dividing the distribution into portions called *strata,* where each stratum would be sampled individually with disproportionately heavy sampling of the more critical strata.

To illustrate, suppose that the distribution is divided into three strata in the manner shown in Table 22.8. These strata were chosen to correspond to observations approximately from 0 to 1, from 1 to 3, and from 3 to , respectively. To ensure that the random observations generated for each stratum actually lie in that portion of the distribution, the uniform random numbers must be converted to the indicated range for *F*(*x*), as shown in the third column of Table 22.8. The number of observations to be generated from each stratum is given in the fourth column.[1] The rightmost column then shows the resulting *sampling weight* for each stratum, i.e., the *ratio* of the *sampling propor-tion* (the fraction of the total sample to be drawn from the stratum) to the *distribution proportion* (the probability of a random observation falling inside the stratum). These sampling weights roughly reflect the relative importance of the respective strata in determining the mean.

Given the formulation of the stratified sampling approach shown in Table 22.8, the same uniform random numbers used in Table 22.7 yield the observations given in the fifth column in Table 22.9. However, it would not be correct to use the unweighted average of these observations to estimate the mean, because certain portions of the distribution have been sampled more than others. Therefore, before we take the average, we divide the observations from each stratum by the sampling weight for that stratum to give proportionate weightings to the different portions of the distribution, as shown in the rightmost column of Table 22.9. The resulting *weighted* average of 0.948 provides the desired estimate of the mean.

[1] These sample sizes are roughly based on a recommended guideline that they be proportional to the *product* of the *probability* of a random observation’s falling inside the corresponding stratum *times* the *standard deviation* within this stratum.

TABLE 22.9 Application of stratified sampling to the example



Stratum	i	Random Number r_i	Stratum Random No. r_i'	Stratum Random Observation $x_i' = -\ln(1 - r_i')$	Sampling Weight w_i	x_i'/w_i
1	1	0.495	0.317	0.381	$\frac{5}{8}$	0.610
	2	0.335	0.215	0.242	$\frac{5}{8}$	0.387
	3	0.791	0.507	0.707	$\frac{5}{8}$	1.131
	4	0.469	0.300	0.357	$\frac{5}{8}$	0.571
2	5	0.279	0.729	1.306	$\frac{5}{4}$	1.045
	6	0.698	0.864	1.995	$\frac{5}{4}$	1.596
	7	0.013	0.644	1.033	$\frac{5}{4}$	0.826
	8	0.761	0.884	2.154	$\frac{5}{4}$	1.723
3	9	0.290	0.9716	3.561	5	0.712
	10	0.693	0.9877	4.398	5	0.880
						Total = 9.481
						Estimate of mean = 0.948

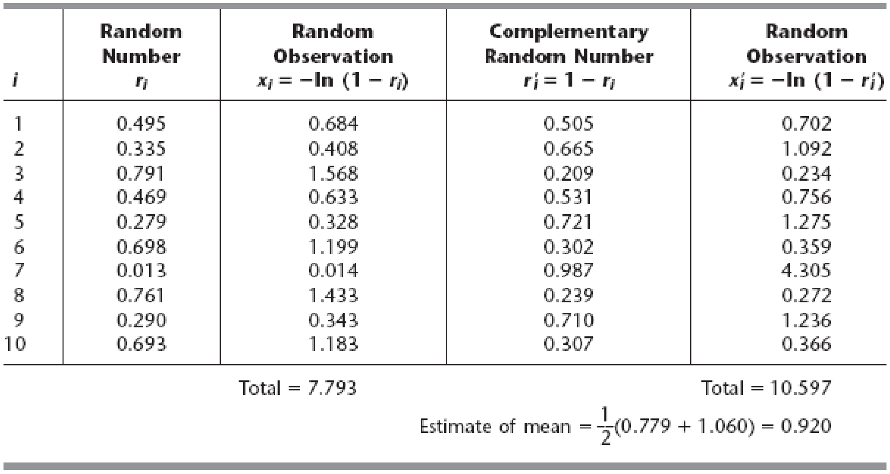
Method of Complementary Random Numbers

The second variance-reducing technique we shall mention is the method of *complemen-tary random numbers.*[1] The motivation for this method is that the "luck of the draw" on the uniform random numbers generated may cause the average of the resulting random observations to be substantially on one side of the true mean, whereas the *complements* of those uniform random numbers (which are themselves uniform random numbers) would have tended to yield a nearly opposite result. (For example, the uniform random numbers in Table 22.7 average less than 0.5, and none are as large as 0.8, which led to an estimate substantially below the true mean.) Therefore, using *both* the original uniform random numbers *and* their complements to generate random observations and then calculating the *combined* sample average should provide a more precise estimator of the mean. This approach is illustrated in Table 22.10,[2] where the first three columns come from Table 22.7 and the two rightmost columns use the complementary uniform random numbers, which results in a combined sample average of 0.920.

[1] This method is a special case of the method of *antithetic variates,* which attempts to generate *pairs* of random observations having a high *negative* correlation, so that the combined average will tend to be closer to the mean.

[2] Note that 20 calculations of a logarithm were required in this case, in contrast to the 10 that were required by each of the preceding techniques.

TABLE 22.10 Application of the method of complementary random numbers to the example



i	Random Number r_i	Random Observation $x_i = -\ln(1 - r_i)$	Complementary Random Number $r_i' = 1 - r_i$	Random Observation $x_i' = -\ln(1 - r_i')$
1	0.495	0.684	0.505	0.702
2	0.335	0.408	0.665	1.092
3	0.791	1.568	0.209	0.234
4	0.469	0.633	0.531	0.756
5	0.279	0.328	0.721	1.275
6	0.698	1.199	0.302	0.359
7	0.013	0.014	0.987	4.305
8	0.761	1.433	0.239	0.272
9	0.290	0.343	0.710	1.236
10	0.693	1.183	0.307	0.366
		Total = 7.793		Total = 10.597

$$\text{Estimate of mean} = \frac{1}{2}(0.779 + 1.060) = 0.920$$

Conclusions

This example has suggested that the variance-reducing techniques provide a much more precise estimator of the mean than does straightforward simulation (the crude Monte Carlo technique). These results definitely were not a coincidence, as a derivation of the variance of the estimators would show. In comparison with straightforward simulation, these techniques (including several more complicated ones not presented here) do indeed provide a much more precise estimator with the same amount of computer time, or they provide an equally precise estimator with much less computer time. Despite the fact that additional analysis may be required to incorporate one or more of these techniques into the simulation study, the rewards should not be forgone readily.

Although this example was particularly simple, it is often possible, though more difficult, to apply these techniques to much more complex problems. For example, suppose that the objective of the simulation study is to estimate the expected waiting time of customers in a queueing system (such as those described in Sec. 18.1). Because both the probability distribution of interarrival times and the probability distribution of service times are involved, and because consecutive waiting times are not statistically independent, this problem may appear to be beyond the capabilities of the variance-reducing techniques. However, as has been described in detail elsewhere,[1] these techniques and others can indeed be applied to this type of problem very advantageously. For example, the method of *complementary random numbers* can be applied simply by repeating the original simulation run, substituting the complements of the original uniform random numbers to generate the corresponding random observations.

[1] S. Ehrenfeld and S. Ben-Tuvia, “The Efficiency of Statistical Simulation Procedures,” *Technometrics,* 4(2): 257–275, 1962. Also see Chap. 11 of Selected Reference 11. For additional information on variance-reducing techniques, see the November 1989 issue of *Management Science* for a special issue on this topic.

22.8 REGENERATIVE METHOD OF STATISTICAL ANALYSIS

The statistical analysis of a simulation run involves using the output to obtain both a point estimate and confidence interval of some steady-state measure (or measures) of performance of the system. (For example, one such measure for a queueing system would be the mean of the steady-state distribution of waiting times for the customers.) To do this analysis, the simulation run can be viewed as a statistical experiment that is generating a series of sample observations of the measure. The question is how to use these sample observations to compute the point estimate and confidence interval.

Traditional Methods and Their Shortcomings

The most straightforward approach would be to use standard statistical procedures to compute these quantities from the observations. However, there are two special characteristics of the observations from a simulation run that require some modification of this approach.

One characteristic is that the system is not in a steady-state condition when the simulation run begins, so the initial observations are not random observations from the underlying probability distribution for the steady-state measure of performance. The traditional approach to circumventing this difficulty is to not start collecting data until it is believed that the simulated system has essentially reached a steady-state condition. Unfortunately, it is difficult to estimate just how long this warm-up period needs to be. Furthermore, available analytical results suggest that a surprisingly long period is required, so that a great deal of unproductive computer time must be expended.

The second special characteristic of a simulated experiment is that its observations are likely to be highly correlated. This is the case, for example, for the waiting times of successive customers in a queueing system. On the other hand, standard statistical procedures for computing the confidence interval for some measure of performance assume that the sample observations are statistically independent random observations from the underlying probability distribution for the measure.

One traditional method of circumventing this difficulty is to execute a series of completely separate and independent simulation runs of equal length and to use the average measure of performance for each run (excluding the initial warm-up period) as an individual observation. The main disadvantage is that each run requires an initial warm-up period for approaching a steady-state condition, so that much of the simulation time is unproductive. The second traditional method eliminates this disadvantage by making the runs consecutively, using the ending condition of one run as the steady-state starting condition for the next run. In other words, one continuous overall simulation run (except for the one initial warm-up period) is divided for bookkeeping purposes into a series of equal portions (referred to as batches). The average measure of performance for each batch is then treated as an individual observation. The disadvantage of this method is that it does not eliminate the correlation between observations entirely, even though it may reduce it considerably by making the portions sufficiently long.

The Regenerative Method Approach

We now turn to an innovative statistical approach that is specially designed to eliminate the shortcomings of the traditional methods described above. (This is the approach used by *Queueing Simulator* to obtain its point estimates and confidence intervals.)

The basic concept underlying this approach is that for many systems a simulation run can be divided into a series of cycles such that the evolution of the system in a cycle is a probabilistic replica of the evolution in any other cycle. Thus, if we calculate an appropriate measure of the length of the cycle along with some *statistic* to summarize the behavior of interest within each cycle, these statistics for the respective cycles constitute a series of independent and identically distributed observations that can be analyzed by standard statistical procedures. Because the system keeps going through these independent and identically distributed cycles regardless of whether it is in a steady-state condition, these observations are directly applicable from the outset for estimating the steady-state behavior of the system.

For cycles to possess these properties, they must each *begin* at the same regeneration point, i.e., at the point where the system probabilistically restarts and can proceed without any knowledge of its past history. The system can be viewed as *regenerating* itself at this point in the sense that the probabilistic structure of the future behavior of the system depends upon being at this point and not on anything that happened previously. (This property is the *Markovian property* described in Sec. 16.2 for Markov chains.) A cycle *ends* when the system again reaches the regeneration point (when the next cycle begins). Thus, the length of a cycle is the elapsed time between consecutive occurrences of the regeneration point. This elapsed time is a random variable that depends upon the evolution of the system.

When *next-event incrementing* is used, a typical regeneration point is a point at which an event has just occurred but no future events have yet been scheduled. Thus, nothing needs to be known about the history of previous schedulings, and the simulation can start from scratch in scheduling future events. When *fixed-time incrementing* is used, a regeneration point is a point at which the probabilities of possible events occurring during the next unit of time do not depend upon when any past events occurred, only on the current state of the system.

Not every system possesses regeneration points, so this regenerative method of collecting data cannot always be used. Furthermore, even when there are regeneration points, the one chosen to define the beginning and ending points of the cycles must recur frequently enough that a substantial number of cycles will be obtained with a reasonable amount of computer time.[1] Thus, some care must be taken to choose a suitable regeneration point.

Perhaps the most important application of the regenerative method to date has been the simulation of queueing systems, including queueing networks (see Sec. 17.9) such as the ones that arise in computer modeling.[2]

Example. Suppose that information needs to be obtained about the steady-state behavior of a system that can be formulated as a *single-server queueing system* (see Sec. 17.2). However, both the interarrival and service times have a *discrete uniform distribution* with a probability of $\frac{1}{10}$ of the values of 6,8,. . . , 24 and the values of 1, 3,. . . ,19,respectively. Because analytical results are not available, simulation with *next-event increment-ing* is to be used to obtain the desired results.

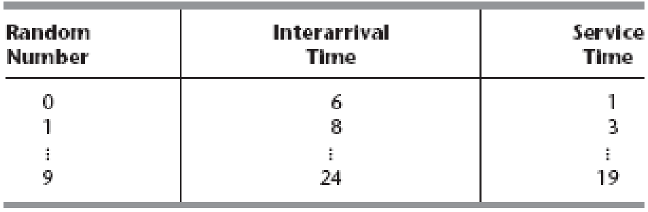
[1] The basic theoretical requirements for the method are that the expected cycle length be *finite* and that the number of cycles would go to infinity if the system continued operating indefinitely. For details, see P. W. Glynn and D. L. Iglehart, "Conditions for the Applicability of the Regenerative Method," *Management Science,* 39: 1108–1111, 1993.

[2] See, e.g., D. L. Iglehart and G. S. Shedler, *Regenerative Simulation of Passage Times in Networks of Queues,*Lecture Notes in Control and Information Sciences, vol. 4, Springer-Verlag, New York, 1980. For another exposition that emphasizes applications to computer system modeling, see G. S. Shedler, *Regeneration and Net-works of Queues,* Springer-Verlag, New York, 1987.

Except for the distributions involved, the general approach is the same as that described in Sec. 22.1 for Example 2. In particular, the building blocks of the simulation model are the same as specified there, including defining the state of the system as the number of customers in the system. Suppose that one-digit random integer numbers are used to generate the random observations from the distributions, as shown in Table 22.11. Beginning the simulation run with no customers in the system then yields the results summarized in Table 22.12 and Fig. 22.16, where the random numbers are obtained sequentially as needed from the tenth row of Table 22.3.[1] (Note in Table 22.12 that, at time 98, the arrival of one customer and the service completion for another customer occur simultaneously, so these canceling events are not visible in Fig. 22.16.)

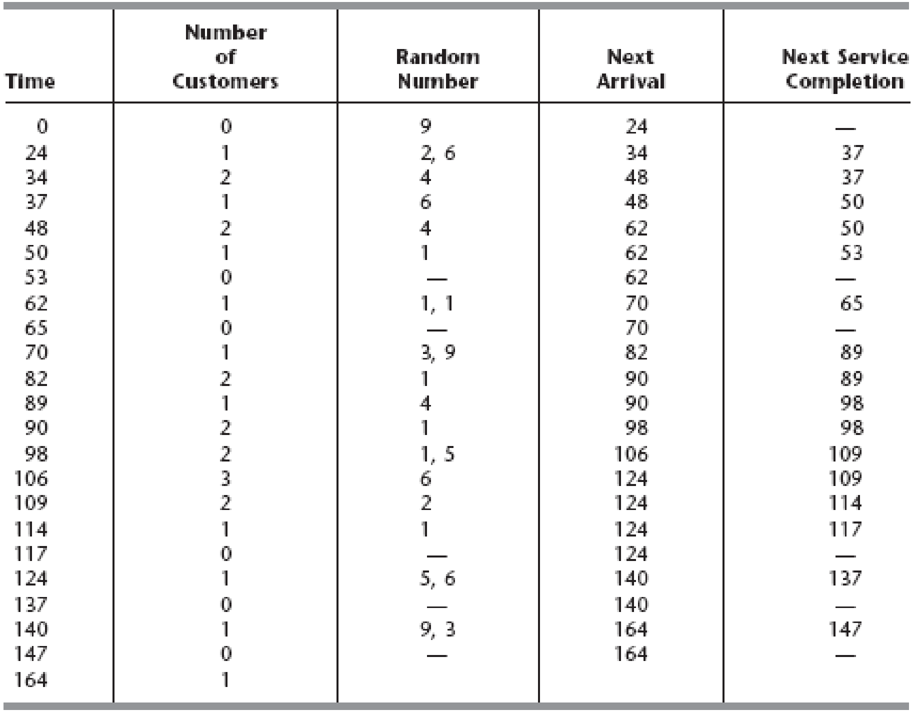
[1] When both an interarrival time and a service time need to be generated at the same time, the interarrival time is obtained first.

TABLE 22.11 Correspondence between random numbers and random observations for the queueing system example



Random Number	Interarrival Time	Service Time
0	6	1
1	8	3
⋮	⋮	⋮
9	24	19

TABLE 22.12 Simulation run for the queueing system example



Time	Number of Customers	Random Number	Next Arrival	Next Service Completion
0	0	9	24	—
24	1	2, 6	34	37
34	2	4	48	37
37	1	6	48	50
48	2	4	62	50
50	1	1	62	53
53	0	—	62	—
62	1	1, 1	70	65
65	0	—	70	—
70	1	3, 9	82	89
82	2	1	90	89
89	1	4	90	98
90	2	1	98	98
98	2	1, 5	106	109
106	3	6	124	109
109	2	2	124	114
114	1	1	124	117
117	0	—	124	—
124	1	5, 6	140	137
137	0	—	140	—
140	1	9, 3	164	147
147	0	—	164	—
164	1			

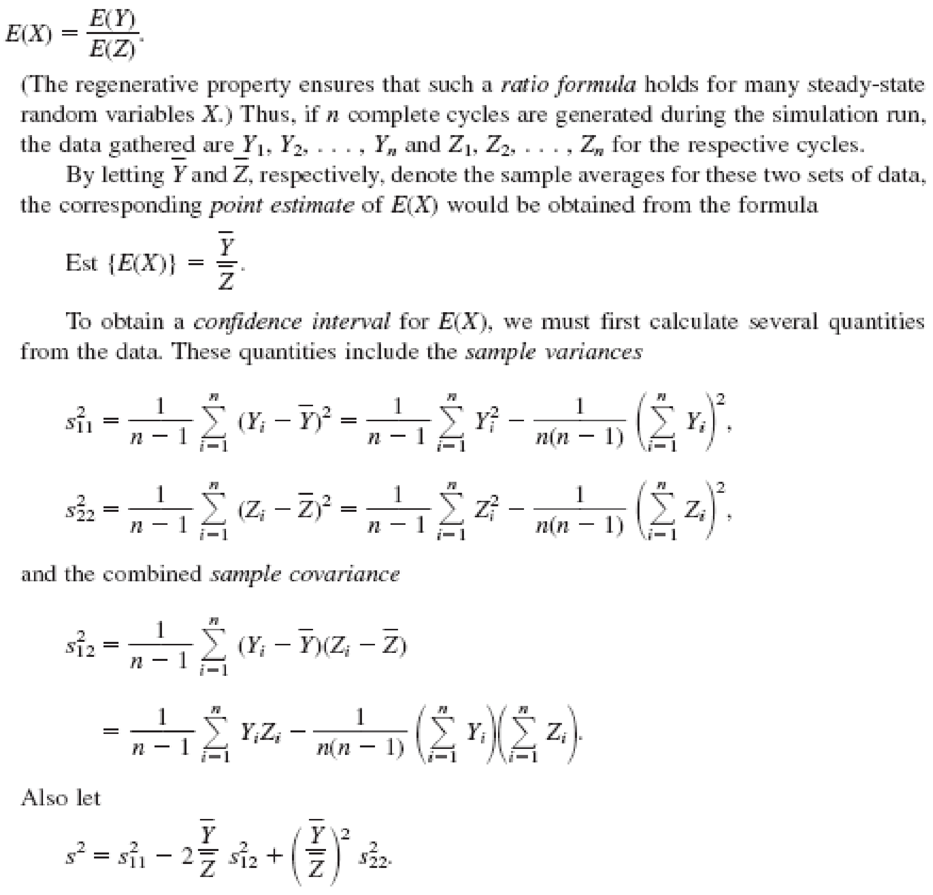
For this system, one *regeneration point* is where an *arrival* occurs with *no* previous customers left. At this point, the process probabilistically restarts, so the probabilistic structure of when future arrivals and service completions will occur is completely independent of any previous history. The only relevant information is that the system has just entered the special state of having had no customers *and* having the time until the next arrival reach zero. The simulation run would not previously have scheduled any future events but would now generate *both* the next interarrival time and the service time for the customer that just arrived.

The only other regeneration points for this system are where an arrival and a service completion occur simultaneously, with a prespecified number of customers in the system. However, the regeneration point described in the preceding paragraph occurs much more frequently and thus is a better choice for defining a cycle. With this selection, the first five complete cycles of the simulation run are those shown in Fig. 22.16. (In most cases, you should have a considerably larger number of cycles in the entire simulation run in order to have sufficient precision in the statistical analysis.)

Various types of information about the steady-state behavior of the system can be obtained from this simulation run, including *point estimates* and *confidence intervals* for the expected number of customers in the system, the expected waiting time, and so on. In each case, it is necessary to use only the corresponding statistics from the respective cycles and the lengths of the cycles. We shall first present the general statistical expressions for the regenerative method and then apply them to this example.

Statistical Formulas

Formally speaking, the statistical problem for the regenerative method is to obtain estimates of the expected value of some random variable *X* of interest. This estimate is to be obtained by calculating a statistic *Y* for each cycle and an appropriate measure *Z* of the *size* of the cycle such that



$$E(X) = \frac{E(Y)}{E(Z)}.$$

(The regenerative property ensures that such a *ratio formula* holds for many steady-state random variables X.) Thus, if n complete cycles are generated during the simulation run, the data gathered are $Y_1, Y_2, \ldots, Y_n$ and $Z_1, Z_2, \ldots, Z_n$ for the respective cycles.

By letting $\overline{Y}$ and $\overline{Z}$, respectively, denote the sample averages for these two sets of data, the corresponding *point estimate* of $E(X)$ would be obtained from the formula

$$\text{Est}\ \{E(X)\} = \frac{\overline{Y}}{\overline{Z}}.$$

To obtain a *confidence interval* for $E(X)$, we must first calculate several quantities from the data. These quantities include the *sample variances*

$$s_{11}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} Y_i\right)^2,$$

$$s_{22}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2 = \frac{1}{n-1}\sum_{i=1}^{n} Z_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} Z_i\right)^2,$$

and the combined *sample covariance*

$$s_{12}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z})$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} Y_i Z_i - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} Z_i\right).$$

Also let

$$s^2 = s_{11}^2 - 2\frac{\overline{Y}}{\overline{Z}}\, s_{12}^2 + \left(\frac{\overline{Y}}{\overline{Z}}\right)^2 s_{22}^2.$$

(

Finally, let α be the constant such that $1 - 2\alpha$ is the desired *confidence coefficient* for the confidence interval, and look up K_α in Table A5.1 (see App. 5) for the normal distribution. If n is not too small, an *asymptotic confidence interval* for $E(X)$ is then given by

$$\frac{\overline{Y}}{\overline{Z}} - \frac{K_\alpha\, s}{\overline{Z}\sqrt{n}} \le E(X) \le \frac{\overline{Y}}{\overline{Z}} + \frac{K_\alpha\, s}{\overline{Z}\sqrt{n}};$$

i.e., the probability is approximately $1 - 2\alpha$ that the endpoints of an interval generated in this way will surround the actual value of $E(X)$.

Application of the Statistical Formulas to the Example

Consider first how to estimate the *expected waiting time* for a customer *before* beginning service (denoted by *Wq* in Chap. 17). Thus, the random variable *X* now represents a customer's waiting time excluding service, so that

$$W_q = E(X).$$

The corresponding information gathered during the simulation run is the *actual* waiting time (excluding service) incurred by the respective customers. Therefore, for each cycle, the summary statistic *Y* is the *sum of the waiting times,* and the size of the cycle *Z* is the *number of customers,* so that

$$W_q = \frac{E(Y)}{E(Z)}.$$

Refer to Fig. 22.16 and Table 22.12; for cycle 1, a total of three customers are processed, so *Z*1 3. The first customer incurs no waiting before beginning service, the second waits 3 units of time (from 34 to 37), and the third waits 2 units of time (from 48 to 50), so *Y*1 5. We proceed similarly for the other cycles. The data for the problem are

$$\begin{aligned}
Y_1 &= 5, & Z_1 &= 3\\
Y_2 &= 0, & Z_2 &= 1\\
Y_3 &= 34, & Z_3 &= 5\\
Y_4 &= 0, & Z_4 &= 1\\
Y_5 &= 0, & Z_5 &= 1\\
\overline{Y} &= 7.8, & \overline{Z} &= 2.2.
\end{aligned}$$

Therefore, the *point estimate* of *Wq* is

$$\text{Est } \{W_q\} = \frac{\overline{Y}}{\overline{Z}} = \frac{7.8}{2.2} = 3\frac{6}{11}.$$

To obtain a 95 percent confidence interval for *Wq*, the preceding formulas are first used to calculate

$$s_{11}^2 = 219.20, \qquad s_{22}^2 = 3.20, \qquad s_{12}^2 = 24.80, \qquad s = 9.14.$$

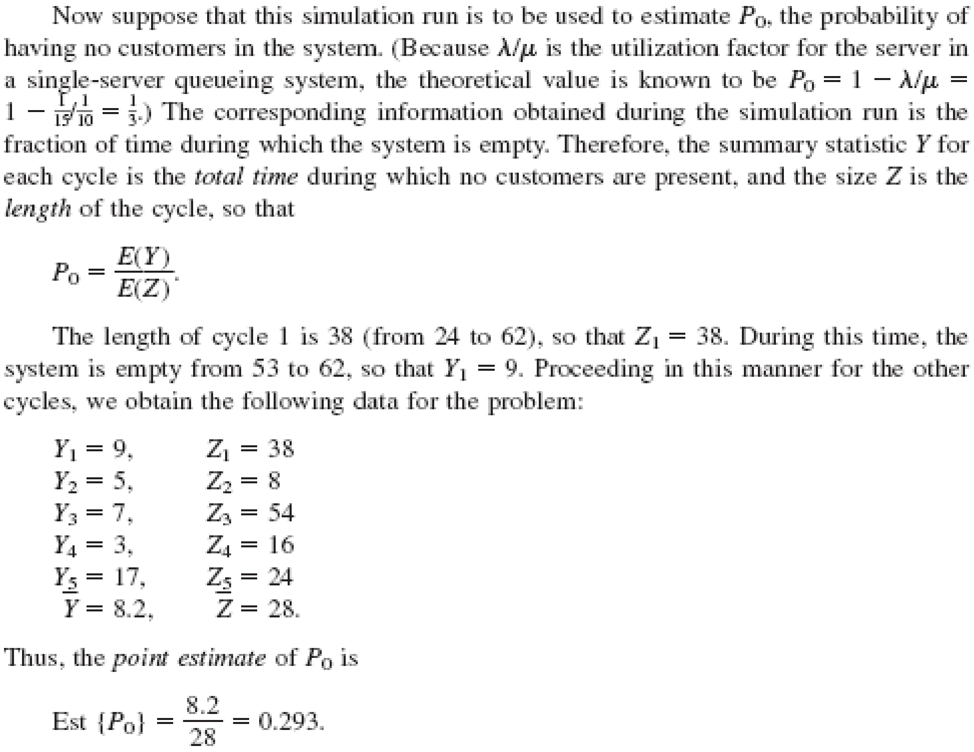
Because $1 - 2\alpha = 0.95$, then $\alpha = 0.025$, so that $K_\alpha = 1.96$ from Table A5.1. The resulting confidence interval is

$$-0.09 \le W_q \le 7.19;$$

or

$$W_q \le 7.19.$$

The reason that this confidence interval is so wide (even including impossible negative values) is that the number of sample observations (cycles), *n* 5, is so small. Note in the general formula that the width of the confidence interval is *inversely pro-portional* to the *square root* of *n*, so that, e.g., quadrupling *n* reduces the width by half (assuming no change in *s* or *Z*). Given preliminary values of *s* and *Z* from a short preliminary simulation run (such as the run in Table 22.12), this relationship makes it possible to estimate in advance the width of the confidence interval that would result from any given choice of *n* for the full simulation run. The final choice of *n* can then be made based on the trade-off between computer time and the precision of the statistical analysis.



Now suppose that this simulation run is to be used to estimate P_0, the probability of having no customers in the system. (Because λ/μ is the utilization factor for the server in a single-server queueing system, the theoretical value is known to be $P_0 = 1 - \lambda/\mu = 1 - \frac{1}{15}/\frac{1}{10} = \frac{1}{3}$.) The corresponding information obtained during the simulation run is the fraction of time during which the system is empty. Therefore, the summary statistic Y for each cycle is the *total time* during which no customers are present, and the size Z is the *length* of the cycle, so that

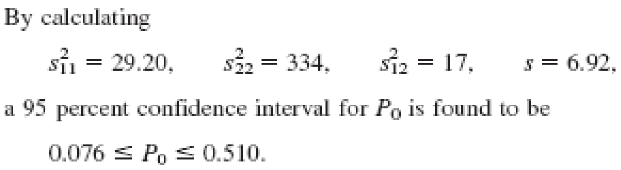
$$P_0 = \frac{E(Y)}{E(Z)}.$$

The length of cycle 1 is 38 (from 24 to 62), so that $Z_1 = 38$. During this time, the system is empty from 53 to 62, so that $Y_1 = 9$. Proceeding in this manner for the other cycles, we obtain the following data for the problem:

$$\begin{aligned} Y_1 &= 9, & Z_1 &= 38 \\ Y_2 &= 5, & Z_2 &= 8 \\ Y_3 &= 7, & Z_3 &= 54 \\ Y_4 &= 3, & Z_4 &= 16 \\ Y_5 &= 17, & Z_5 &= 24 \\ \bar{Y} &= 8.2, & \bar{Z} &= 28. \end{aligned}$$

Thus, the *point estimate* of P_0 is

$$\text{Est}\ \{P_0\} = \frac{8.2}{28} = 0.293.$$



By calculating

$$s_{11}^2 = 29.20, \qquad s_{22}^2 = 334, \qquad s_{12}^2 = 17, \qquad s = 6.92,$$

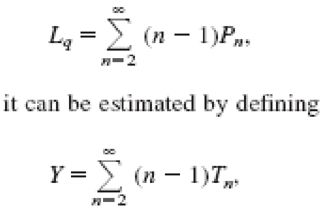
a 95 percent confidence interval for P_0 is found to be

$$0.076 \le P_0 \le 0.510.$$

(The wide range of this interval indicates that a much longer simulation run would be needed to obtain a relatively precise estimate of *P*0.)

If we redefine *Y* appropriately, the same approach also can be used to estimate other probabilities involving the number of customers in the system. However, because this number never exceeded 3 during this simulation run, a much longer run will be needed if the probability involves larger numbers.

The other basic expected values of queueing theory defined in Sec. 17.2 (W, Lq, and *L*) can be estimated from the estimate of *Wq* by using the relationships among these four expected values given near the end of Sec. 17.2. However, the other expected values can also be estimated directly from the results of the simulation run. For example, because the expected number of customers waiting to be served is

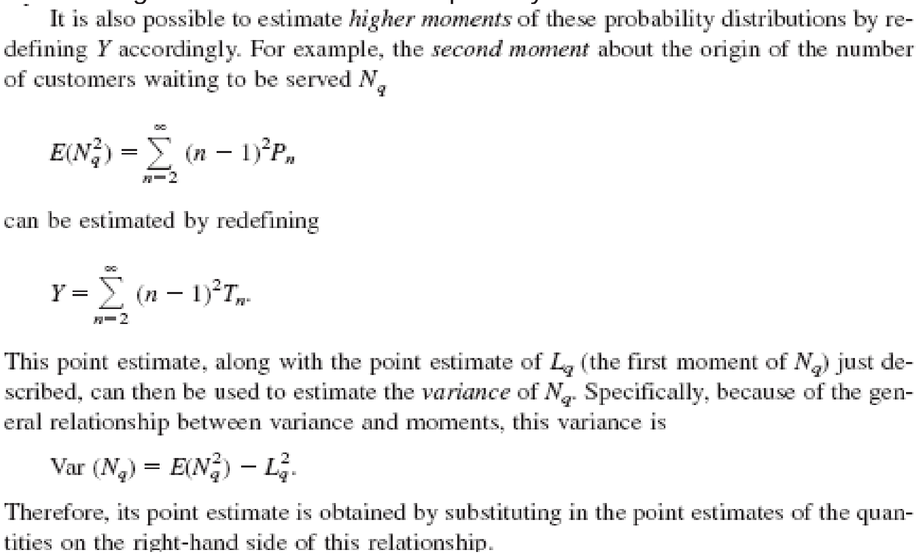


$$L_q = \sum_{n=2}^{\infty} (n-1)P_n,$$

it can be estimated by defining

$$Y = \sum_{n=2}^{\infty} (n-1)T_n,$$

where *Tn* is the *total time* that exactly *n* customers are in the system during the cycle. (This definition of *Y* actually is equivalent to the definition used for estimating *Wq*.) In this case, *Z* is defined as it would be for estimating any *Pn*, namely, the *length* of the cycle. The resulting *point estimate* of *Lq* then turns out to be simply the *point estimate* of *Wq multiplied by* the actual *average arrival rate* for the complete cycles observed.



It is also possible to estimate *higher moments* of these probability distributions by redefining Y accordingly. For example, the *second moment* about the origin of the number of customers waiting to be served N_q

$$E(N_q^2) = \sum_{n=2}^{\infty} (n-1)^2 P_n$$

can be estimated by redefining

$$Y = \sum_{n=2}^{\infty} (n-1)^2 T_n.$$

This point estimate, along with the point estimate of L_q (the first moment of N_q) just described, can then be used to estimate the *variance* of N_q. Specifically, because of the general relationship between variance and moments, this variance is

$$\text{Var}\,(N_q) = E(N_q^2) - L_q^2.$$

Therefore, its point estimate is obtained by substituting in the point estimates of the quantities on the right-hand side of this relationship.

Therefore, its point estimate is obtained by substituting in the point estimates of the quantities on the right-hand side of this relationship.

Finally, we should mention that it was unnecessary to generate the first *interarrival* time (24) for the simulation run summarized in Table 22.12 and Fig. 22.16, because this time played no role in the statistical analysis. It is more efficient with the regenerative method just to start the run at the regeneration point.

Selected Reference 5 provides considerably more information about the regenerative method, including how it can be applied to more complicated kinds of problems than those considered here. (Also see the references given in the second footnote at the beginning of this section.)

22.9 CONCLUSIONS

Simulation is a widely used tool for estimating the performance of complex stochastic systems if contemplated designs or operating policies are to be used.

We have focused in this chapter on the use of simulation for predicting the *steady-state* behavior of systems whose states change only at discrete points in time. However, by having a series of runs begin with the prescribed *starting conditions,* we can also use simulation to describe the *transient* behavior of a proposed system. Furthermore, if we use differential equations, simulation can be applied to systems whose states change *con-tinuously* with time.

Simulation is one of the most popular techniques of operations research because it is such a flexible, powerful, and intuitive tool. In a matter of seconds or minutes, it can simulate even years of operation of a typical system while generating a series of statistical observations about the performance of the system over this period. Because of its exceptional versatility, simulation has been applied to a wide variety of areas. Furthermore, its horizons continue to broaden because of the great progress being made in simulation software, including software for performing simulations on spreadsheets.

On the other hand, simulation should not be viewed as a panacea when studying stochastic systems. When applicable, analytical methods (such as those presented in Chaps. 15 to 21) have some significant advantages. Simulation is inherently an imprecise technique. It provides only *statistical estimates* rather than exact results, and it *compares al-ternatives* rather than generating an optimal one. Furthermore, despite impressive advances in software, simulation still can be a relatively *slow and costly* way to study complex stochastic systems. For such systems, it usually requires a large amount of time and expense for analysis and programming, in addition to considerable computer running time. Simulation models tend to become unwieldy, so that the number of cases that can be run and the accuracy of the results obtained often turn out to be inadequate. Finally, simulation yields only *numerical data* about the performance of the system, so that it provides no additional insight into the cause-and-effect relationships within the system except for the clues that can be gleaned from these numbers (and from the analysis required to construct the simulation model). Therefore, it is very expensive to conduct a sensitivity analysis of the parameter values assumed by the model. The only possible way would be to conduct new series of simulation runs with different parameter values, which would tend to provide relatively little information at a relatively high cost.

For all these reasons, analytical methods (when available) and simulation have important complementary roles for studying stochastic systems. An analytical method is well suited for doing at least preliminary analysis, for examining cause-and-effect relationships, for doing some rough optimization, and for conducting sensitivity analysis. When the mathematical model for the analytical method does not capture all the important features of the stochastic system, simulation is well suited for incorporating all these features and then obtaining detailed information about the measures of performance of the few leading candidates for the final system configuration.

Simulation provides a way of *experimenting* with proposed systems or policies without actually implementing them. Sound statistical theory should be used in designing these experiments. Surprisingly long simulation runs often are needed to obtain *statistically sig-nificant* results. However, *variance-reducing techniques* can be very helpful in reducing the length of the runs needed.

Several tactical problems arise when we apply traditional statistical estimation procedures to simulated experiments. These problems include prescribing appropriate *start-ing conditions,* determining how long a *warm-up period* is needed to essentially reach a steady-state condition, and dealing with *statistically dependent* observations. These problems can be eliminated by using the *regenerative method* of statistical analysis. However, there are some restrictions on when this method can be applied.

Simulation unquestionably has a very important place in the theory and practice of OR. It is an invaluable tool for use on those problems where analytical techniques are inadequate, and its usage is continuing to grow.