SUPPLEMENT 2 TO CHAPTER 20

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 the *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* mentioned at the beginning of Chap. 19 and described in detail in Sec. 29.2 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.

[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.

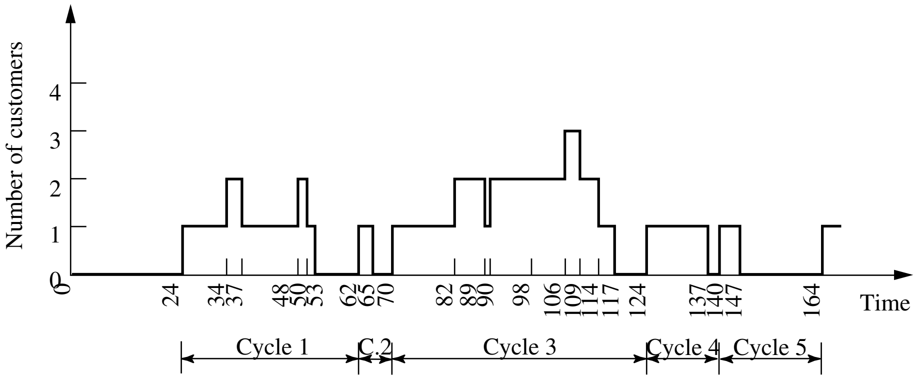

FIGURE 1 Outcome of the simulation run for the queueing system example.

TABLE 1 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

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 incrementing* is to be used to obtain the desired results.

Except for the distributions involved, the general approach is the same as that described in Sec. 20.1 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 1. Beginning the simulation run with no customers in the system then yields the results summarized in Table 2 and Fig. 1, where the random numbers are obtained sequentially as needed from the tenth row of Table 20.3.[3] (Note in Table 2 that, at time 98, the arrival of

[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 Networks of Queues,* Springer-Verlag, New York, 1987. For a general introduction to the regenerative method that describes how it can also be applied to more complicated kinds of problems than those considered here, see M. A. Crane and A. J. Lemoine, *An Introduction to the Regenerative Method for Simulation Analysis,* Springer-Verlag, Berlin, 1977.

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

TABLE 2 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			

one customer and the service completion for another customer occur simultaneously, so these canceling events are not visible in Fig. 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. 1. (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*

$$\begin{aligned} s_{12}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})(Z_i - \overline{Z}) \\ &= \frac{1}{n-1} \sum_{i=1}^{n} Y_i Z_i - \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} Y_i \right)\left(\sum_{i=1}^{n} Z_i \right). \end{aligned}$$

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 W_q 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. 1 and Table 2; 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

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

Therefore, the *point estimate* of W_q 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 W_q, 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 proportional* to the *square root* of n, so that, e.g., quadrupling n reduces the width by half (assuming no change in s or $\overline{Z}$). Given preliminary values of s and $\overline{Z}$ from a short preliminary simulation run (such as the run in Table 2), 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, L_q, and L) can be estimated from the estimate of W_q 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 T_n 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 W_q.) In this case, Z is defined as it would be for estimating any P_n, namely, the *length* of the cycle. The resulting *point estimate* of L_q then turns out to be simply the *point estimate* of W_q *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.

Finally, we should mention that it was unnecessary to generate the first *interarrival* time (24) for the simulation run summarized in Table 2 and Fig. 1, 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.

PROBLEMS

20S2-1. A certain single-server system has been simulated, with the following sequence of waiting times before service for the respective customers. Use the regenerative method to obtain a point estimate and 90 percent confidence interval for the steady-state expected waiting time before service.

(a) 0, 5, 4, 0, 2, 0, 3, 1, 6, 0

(b) 0, 3, 2, 0, 3, 1, 5, 0, 0, 2, 4, 0, 3, 5, 2, 0

20S2-2. Consider the queueing system example presented in this supplement for the regenerative method. Explain why the point where a *service completion* occurs with *no* other customers left is *not* a regeneration point.

20S2-3. Reconsider Prob. 20.6-3. You now wish to begin the analysis by performing a short simulation by hand and then applying the regenerative method of statistical analysis when possible.

R **(a)** Starting with four new relays, simulate the operation of the two alternative policies for 5,000 hours of simulated time. Obtain the needed uniform random numbers as instructed at the beginning of the Problems section for Chap. 20.

(b) Use the data from part (*a*) to make a preliminary comparison of the two alternatives on a cost basis.

(c) For the *proposed* policy, describe an appropriate regeneration point for defining cycles that will permit applying the regenerative method of statistical analysis. Explain why the regenerative method cannot be applied to the *current* policy.

(d) For the proposed policy, use the regenerative method to obtain a point estimate and 95 percent confidence interval for the steady-state expected cost per hour from the data obtained in part (*a*).

(e) Write a computer simulation program for the two alternative policies. Then repeat parts (*a*), (*b*), and (*d*) on the computer, with 100 cycles for the proposed policy and 55,000 hours of simulated time (including a warm-up period of 5,000 hours) for the current policy.

20S2-4. One of the main lessons of queueing theory (Chap. 17) is that the amount of variability in the service times and interarrival times has a substantial impact on the measures of performance of the queueing system. Significantly decreasing variability helps considerably.

This phenomenon is well illustrated by the *M*/*G*/1 queueing model presented at the beginning of Sec. 17.7. For this model, the four fundamental measures of performance (L, L_q, W, and W_q) are expressed directly in terms of the *variance* of service times (σ^2), so we can see immediately what the impact of decreasing σ^2 would be.

Consider an *M*/*G*/1 queueing system with mean arrival rate $\lambda = 0.8$ and mean service rate $\mu = 1$, so the utilization factor is $\rho = \lambda/\mu = 0.8$.

Q **(a)** Use the Queueing Simulator to execute a simulation run with 10,000 customer arrivals for each of the following cases: (i) $\sigma = 1$ (corresponds to an exponential distribution of service times), (ii) $\sigma = 0.5$ (corresponds to an Erlang distribution of service times with shape parameter $k = 4$), and (iii) $\sigma = 0$ (constant service times). Using the point estimates of L_q obtained, calculate the ratio of L_q for case (ii) to L_q for case (i). Also calculate the ratio of L_q for case (iii) to L_q for case (i).

(b) For each of the three cases considered in part (*a*), use the formulas given in Sec. 17.7 to compute the exact values of L, L_q, W, and W_q. Compare these exact values to the point estimates and 95 percent confidence intervals obtained in part (*a*). Identify any exact values that fall outside the 95 percent confidence interval. Also calculate the exact values of the ratios requested in part (*a*).

20S2-5. Follow the instructions of part (*a*) of Prob. 20S2-4 for an *M*/*G*/2 queueing system (two servers), with $\lambda = 1.6$ and $\mu = 1$ [so $\rho = \lambda/(2\mu) = 0.8$] and with σ^2 still being the variance of service times.

20S2-6. Reconsider Prob. 20S2-4. For the single-server queueing system under consideration, suppose now that service times definitely have an exponential distribution. However, it now is possible to reduce the variability of *interarrival times,* so we want to explore the impact of doing so.

Assume now that $\lambda = 1$ and $\mu = 1.25$, so $\rho = 0.8$. Let σ^2 now denote the variance of interarrival times.

Follow the instructions of Prob. 20S2-4*a*, where the distributions for the three cases now are for interarrival times instead of service times.