S. S. Lavenberg
D. R. Slutz

Introduction to Regenerative Simulation

Abstract: A recently developed method for estimating confidence intervals when simulating stochastic systems having a regenerative structure is reviewed. The paper is basically tutorial, but also considers the pragmatic issue of the simulation duration required to obtain valid estimates. The method is illustrated in terms of simulating the M/G/1 queue. Analytic results for the M/G/1 queue are used to determine the validity of the simulation results.

Introduction

The simulation of a stochastic system such as a queuing system is a statistical experiment. In order to draw meaningful conclusions from such an experiment it is necessary to make statistically valid statements about the outcomes of the experiment. Suppose, for example, a queuing system is simulated in order to estimate a response variable Q (e.g., the long-run average time spent queuing for service). In addition to obtaining a point estimate $\hat{Q}$ of Q, it is desirable to estimate a confidence interval for Q. An estimated *$100 \cdot \alpha\%$ confidence interval* for Q is an interval $(\hat{Q}_1, \hat{Q}_2)$ whose endpoints $\hat{Q}_1$ and $\hat{Q}_2$ are estimated via simulation and have the property that $\Pr\{\hat{Q}_1 < Q < \hat{Q}_2\} = \alpha$. (Note that $\hat{Q}$, $\hat{Q}_1$ and $\hat{Q}_2$ are random variables while Q is a number.) Thus, an estimated confidence interval carries with it a statement that the response variable is contained in the interval with a given probability.

A number of different methods exist for estimating confidence intervals via simulation [1]. Generally, these methods make use of the exact, approximate, or assumed normality of an estimate. The variance of the estimate is unknown and must also be estimated via simulation. The variance is typically estimated a) from independent and identically distributed (i.i.d.) samples of the estimate obtained from independent replications of the simulation, b) by dividing a single run of the simulation into approximately independent subruns and treating the estimates obtained from the subruns as i.i.d., or c) by analyzing the time series of observations (e.g., waiting times of successive customers, queue lengths sampled at equal time increments) from a single run. In the latter case, initial observations are usually discarded and the remaining time series is assumed to be stationary, i.e., the simulation is assumed to be in the "steady state."

Recently, several methods have been developed for estimating confidence intervals for certain response variables when simulating stochastic systems having a regenerative structure [2–4]. Informally, a stochastic system is said to be *regenerative* if with probability one there exists an infinite sequence of increasing random times, called *regeneration points*, at which the system "stochastically restarts." The evolution in time of the system between successive regeneration points is called a *tour* or cycle and the stochastic behavior of the system during different tours is independent and identical. This underlying regenerative structure guarantees that for many response variables, estimates for the response variables based on a single run of the simulation are approximately normal if the run is sufficiently long and if certain random variables associated with a tour (e.g., the time duration of a tour) have finite first two moments. Furthermore, the variance of the estimates can be estimated either from independent replications of the simulation [4] or by observing a fixed number of tours during a single run of the simulation [2, 3]. Thus, provided that a simulation run is sufficiently long, a theoretical basis exists for estimating confidence intervals for many response variables in regenerative stochastic systems.

This paper contains a tutorial exposition of a method for estimating confidence intervals for response variables when simulating a regenerative stochastic system. The method is illustrated in terms of simulating the M/G/1 queue in order to estimate the average queuing time and the average waiting time. The next section contains a concise mathematical derivation of confidence intervals for the average queuing time and average waiting time for the M/G/1 queue. The third section addresses the pragmatic issue of simulation duration. For the M/G/1 queue, analytic approximations are given for the expected number of tours that must be simulated to obtain a given confidence interval width. In addition, confidence interval widths and coverages (the coverage of a confidence interval is the probability that the interval contains the response variable being estimated) are estimated by simulating the M/G/1 queue. Based on the simulation results, conclusions are reached about the number of tours that must be simulated to obtain the required coverage and the validity of the analytic approximations on simulation duration is checked. The last section contains concluding comments.

One motivation for this paper is to provide background for a subsequent paper [5] in which the method is applied to the simulation of a more complex queuing system. The particular M/G/1 queue considered corresponds to a special case of the system in [5] which, generally, is not analytically tractable. Other discussions of the application of regenerative simulation methods to queuing systems can be found in the literature [6–8].

Derivation of confidence intervals

Consider an M/G/1 queue, i.e., a single server queue, with Poisson arrivals at rate λ, $0 < \lambda < \infty$, and i.i.d. service times. Each service time is distributed as a nonnegative random variable T having mean $E[T] = \beta$, $0 < \beta < \infty$, and second moment $E[T^2] = \beta_2$, $0 < \beta_2 < \infty$ ($E[\cdot]$ denotes expectation). Let $\rho = \lambda\beta$ denote the *traffic intensity* for the M/G/1 queue, let q_k denote the time spent queuing for service by the kth customer to arrive, and denote by w_k the waiting time for this customer (waiting time equals queuing time plus service time). Then if $\rho < 1$ the average queuing time

$$Q = \lim_{n\to\infty} \sum_{k=1}^{n} q_k/n$$

and the average waiting time

$$W = \lim_{n\to\infty} \sum_{k=1}^{n} w_k/n$$

exist with probability one and are finite. It is well known that $Q = (\lambda/2)(\beta_2/(1-\lambda\beta))$ and $W = Q + \beta$. Suppose, however, that Q or W is to be estimated via simulation. It is next shown how both a point estimate and an estimated confidence interval for Q can be obtained via simulation based on the regenerative structure of the M/G/1 queue. A point estimate and estimated confidence interval for W can easily be obtained from the estimates for Q.

Suppose at time $t_0 = 0$ a customer arrives at the empty system. It is known that if $\rho < 1$, the event of a customer arriving at the empty system occurs infinitely often with probability one as the system evolves in time, and if $\rho > 1$ there is a positive probability this event will never recur. It is assumed in what follows that $\rho < 1$. Let $\{t_k : k = 1, 2, \cdots\}$ denote the infinite sequence of increasing random times at which a customer arrives at the empty system. The stochastic behavior of the system between any two such successive times is independent of and identical to the stochastic behavior of the system between any two other such successive times. Thus, the system stochastically restarts at each time t_k. The t_k are regeneration points and the evolution of the system between t_{k-1} and t_k is the kth tour. (The above discussion is informal and suits the purpose of this paper. For a more rigorous discussion see [2].)

Denote by ν_k the number of customers served during the kth tour and by σ_k the sum of the queuing times for all customers served during the kth tour. Clearly each of the sequences $\{\nu_k : k = 1, 2, \cdots\}$ and $\{\sigma_k : k = 1, 2, \cdots\}$ is a sequence of i.i.d. random variables. For each k, ν_k is distributed as a random variable ν and σ_k is distributed as a random variable σ. It is known since $\rho < 1$ that $E[\nu] < \infty$, $E[\sigma] < \infty$ and, with probability one,

$$Q = \lim_{n\to\infty} \sum_{k=1}^{n} \sigma_k \Big/ \sum_{k=1}^{n} \nu_k = E[\sigma]/E[\nu].$$

This suggests that a *point estimate* of Q can be obtained by observing a fixed number, say n, of tours during a single run of a simulation of the M/G/1 queue and computing

$$Q(n) = \sum_{k=1}^{n} \sigma_k \Big/ \sum_{k=1}^{n} \nu_k.$$

Note that $\lim_{n\to\infty} Q(n) = Q$ with probability one, in which case $Q(n)$ is said to be a strongly consistent estimate of Q.

Consider next the random variable $\sigma_k - Q\nu_k$, which has mean zero. It is known that $\rho < 1$ and $\beta_2 < \infty$ imply that $E[\nu^2] < \infty$. If, in addition, T has a finite fourth moment, i.e., $E[T^4] = \beta_4 < \infty$, then $E[\sigma^2] < \infty$ so that $\sigma_k - Q\nu_k$ has finite variance

$$V = \mathrm{Var}[\sigma - Q\nu]$$
$$= \mathrm{Var}[\sigma] - 2Q\mathrm{Cov}[\sigma, \nu] + Q^2\mathrm{Var}[\nu],$$

where $\mathrm{Var}[\cdot]$ denotes variance and $\mathrm{Cov}[\cdot, \cdot]$ denotes covariance. The central limit theorem [9] can be applied to the sequence $\{\sigma_k - Q\nu_k \pm k = 1, 2, \cdots\}$ of i.i.d., mean zero, and variance V random variables, and yields

$$\lim_{n\to\infty} \Pr\{[\sigma(n) - Q\nu(n)]/(V/n)^{\frac{1}{2}} < t\} = \phi(t), \qquad (1)$$

where

$$\sigma(n) = \sum_{k=1}^{n} \sigma_k/n,$$

$$\nu(n) = \sum_{k=1}^{n} \nu_k/n,$$

Table 1 Analytic results for M/G/1 queue; $T = T_1 + T_2 + T_3$; T_1 and T_3 uniform on [1, 3] and T_2 uniform on [4, 12].

Traffic intensity ρ	*Waiting time* W	*Customers per tour* $E[\nu]$	*Number of tours* $n(0.95, 0.1)$	*Number of customers* $n(0.95, 0.1)$ $E[\nu]$
0.1	12.7	1.11	75	83
0.2	13.6	1.25	196	245
0.3	14.7	1.43	392	560
0.4	16.2	1.67	929	1190
0.5	18.3	2.00	1260	2520
0.6	21.4	2.50	2240	5590
0.7	26.6	3.33	4150	13,800
0.8	37.0	5.00	8550	42,800
0.9	68.3	10.0	23,400	234,000

and $\phi(t) = (1/2\pi)^{\frac{1}{2}} \int_{-\infty}^{t} e^{-x^2/2}\,dx$ is the probability distribution function of a normal random variable having mean zero and variance one. A strongly consistent estimate for V can be computed based on observing n tours as follows:

Let

$$V_1(n) = \sum_{k=1}^{n} [\sigma_k - \sigma(n)]^2/(n-1),$$

$$V_2(n) = \sum_{k=1}^{n} [\nu_k - \nu(n)]^2/(n-1)$$

and

$$V_{12}(n) = \sum_{k=1}^{n} [\sigma_k - \sigma(n)][\nu_k - \nu(n)]/(n-1).$$

Then

$$V(n) = V_1(n) - 2Q(n)V_{12}(n) + [Q(n)]^2 V_2(n)$$

is a strongly consistent estimate of V. It can be shown using Theorem 4.4.8 of Chung [10] that if V in (1) is replaced by a random variable, such as $V(n)$, which depends on n and tends to V with probability one in the limit as $n \to \infty$, then (1) still holds; i.e.,

$$\lim_{n\to\infty} \Pr\{[\sigma(n) - Q\nu(n)]/[V(n)/n]^{\frac{1}{2}} < t\} = \phi(t). \quad (2)$$

It follows from (2) that if $t > 0$,

$$\lim_{n\to\infty} \Pr\Big\{Q(n) - \{t[V(n)/n]^{\frac{1}{2}}/\nu(n)\} < Q < Q(n) + \{t[V(n)/n]^{\frac{1}{2}}/\nu(n)\}\Big\} = 2\phi(t) - 1.$$

Thus, for n sufficiently large,

$$I_Q(n, \alpha) = [Q(n) - \delta(n, \alpha), Q(n) + \delta(n, \alpha)]$$

is approximately a *100 · α% confidence interval* for Q (i.e., the probability that Q is contained in $I_Q(n, \alpha)$ is approximately α), where

$$\delta(n, \alpha) = \phi^{-1}[(1+\alpha)/2][V(n)/n]^{\frac{1}{2}}/\nu(n)$$

and $\phi^{-1}(\cdot)$ is the inverse of the function $\phi(\cdot)$. (If $\alpha = 0.95$, corresponding to a 95% confidence interval, then $\phi^{-1}[(1+\alpha)/2] = 1.960$.) The derivation of this approximate confidence interval makes use of the regenerative structure of the M/G/1 queue. In addition, it is necessary that $E[\nu] < \infty$, $E[\sigma] < \infty$, $E[\nu^2] < \infty$ and $E[\sigma^2] < \infty$, conditions which hold for the M/G/1 queue if $\rho < 1$ and $\beta_4 < \infty$. A point estimate for the average waiting time W is given by

$$W(n) = Q(n) + \beta, \quad (3)$$

and for n sufficiently large,

$$I_W(n, \alpha) = [W(n) - \delta(n, \alpha), W(n) + \delta(n, \alpha)] \quad (4)$$

is approximately a $100 \cdot \alpha\%$ confidence interval for W. (Law [8] shows that for a class of queues which includes the M/G/1 queue, it is more efficient to estimate W by adding the mean service time to an estimate of Q than it is to estimate W directly.)

Simulation duration

In this section results pertaining to simulation duration are presented for an M/G/1 queue whose service time T is equal to the sum of three mutually independent random variables T_1, T_2 and T_3 where T_1 and T_3 are uniformly distributed on the interval [1, 3] and T_2 is uniformly distributed on the interval [4, 12]. This particular choice for T is motivated by a subsequent paper [5] where simulation of a complex queuing system is considered. The queuing system in [5] degenerates to the above M/G/1 queue in the simplest case.

First, analytic results pertaining to the simulation duration required to achieve a given confidence interval width for W are presented. Law [8] has shown that for the M/G/1 queue $V = \mathrm{Var}[\sigma - Q\nu]$ is given by

$$V = \lambda(E[\nu])^3[\lambda\beta_4/4 + (1+\lambda\beta)\beta_3/3 + 5\lambda^2 E[\nu]\beta_3\beta_2/6 + \lambda^3(E[\nu])^2\beta_2^3/2 + \lambda(1+\lambda\beta)E[\nu]\beta_2^2/4]$$

where $\beta_k = E[T^k]$, $k = 2, 3, 4$, and $E[\nu] = 1/(1-\lambda\beta)$. From (1), for n sufficiently large $n^{\frac{1}{2}}[Q(n) - Q]$ is approximately distributed as a normal random variable having mean zero and variance $V/E[\nu]^2$. Thus, for n sufficiently large the width of a $100 \cdot \alpha\%$ confidence interval for Q is approximately $2\Delta(n, \alpha)$ where

$$\Delta(n, \alpha) = \phi^{-1}[(1+\alpha)/2](V/n)^{\frac{1}{2}}/E[\nu]. \quad (5)$$

The quantity $2\Delta(n, \alpha)$ is also the approximate width of a $100 \cdot \alpha\%$ confidence interval for W. It follows that the number of tours required for the width of a $100 \cdot \alpha\%$ confidence interval for W to be equal to $\delta \cdot W$ is approximately

Table 2 Simulation results for M/G/1 queue of Table 1 based on 100 independent replications for each row of the table.

Traffic intensity ρ	*Waiting time* W	*Number of tours* n	*Point estimate* $\overline{W}(n)$	*Approximate theoretical width* $2\Delta(n, 0.95)$	*Estimated width* $2\bar{\delta}(n, 0.95)$	*Estimated coverage* $\bar{c}(n, 0.95)$
0.2	13.6	50	13.6	2.68	2.30	0.83
		100	13.5	1.90	1.62	0.81
		250	13.6	1.20	1.13	0.91
		500	13.6	0.848	0.819	0.93
		1000	13.6	0.600	0.588	0.91
		5000	13.6	0.268	0.268	0.97
0.5	18.3	50	18.2	9.16	6.74	0.79
		100	18.2	6.48	5.04	0.91
		250	18.3	4.10	3.76	0.90
		500	18.1	2.90	2.75	0.91
		1000	18.3	2.05	2.00	0.91
0.8	37.0	50	37.1	48.4	27.2	0.68
		100	37.7	34.2	24.3	0.78
		250	36.3	21.6	16.0	0.79
		500	36.8	15.3	12.7	0.88
		1000	36.5	10.8	9.64	0.87

$$n(\alpha, \delta) = 4\{\phi^{-1}[(1+\alpha)/2]\}^2 V/(\mathrm{E}[\nu]\delta \cdot W)^2.$$

The expected number of customers served during $n(\alpha, \delta)$ tours is $n(\alpha, \delta)\mathrm{E}[\nu]$. Computed values of W, $\mathrm{E}[\nu]$, $n(0.95, 0.1)$ and $n(0.95, 0.1)\mathrm{E}[\nu]$ for various traffic intensities ρ are given in Table 1. (The resulting M/G/1 queue has $\beta = 12$, $\beta_2 = 150$, $\beta_3 = 1944$ and $\beta_4 = 25993.6$.) Note that both the number of tours and the expected number of customers per tour increase rapidly as ρ approaches one. It is easy to show that

$$\lim_{\lambda \to 1/\beta} n(\alpha, \delta) = \left\{8\{\phi^{-1}[(1+\alpha)/2]\}^2 \beta^2 \beta_2/\delta^2\right\} / \lim_{\lambda \to 1/\beta} (1 - \lambda\beta)$$

so that $n(\alpha, \delta)$ increases as $1/(1-\rho)$ and $n(\alpha, \delta)\mathrm{E}[\nu]$ increases as $1/(1-\rho)^2$ as ρ approaches one.

The confidence interval $I_W(n, \alpha)$ for the average waiting time W given by (4) is an approximate $100 \cdot \alpha\%$ confidence interval, i.e., the probability that W is contained in $I_W(n, \alpha)$ is approximately equal to α, with the approximation becoming better as n is increased. For given n the probability that W is contained in $I_W(n, \alpha)$ is called the true coverage. Additionally, the point estimate $W(n)$ given by (3) is, in general, biased for finite n, i.e., $\mathrm{E}[W(n)] \neq W$, but for n sufficiently large the bias is small. A question arises as to how large n must be for the confidence interval approximation to be satisfactory (i.e., for the true coverage to approach α) and for the bias to be small. Since W is known, this question can be empirically investigated by performing a large number, say 100, of independent replications of a simulation; each replication is terminated after n tours and a point estimate and a confidence interval estimate for W are obtained on each replication. The average of $W(n) - W$ over the replications is an estimate of the bias. In addition, the fraction of replications for which the estimated confidence interval contains W is an estimate of the true coverage.

The M/G/1 queue was simulated in order to estimate the bias, true coverage and confidence interval width. Traffic intensities of 0.2, 0.5 and 0.8 were considered and simulation results for different numbers of tours are presented in Table 2. Each row of Table 2 presents results obtained from 100 independent replications of a simulation where each replication was terminated after n tours had been completed. In the table $\overline{W}(n)$ and $2\bar{\delta}(n, 0.95)$ denote the averages over the 100 replications of, respectively, the point estimate of W and the width of the estimated 95% confidence interval for W. The quantity $2\Delta(n, 0.95)$ is the approximate theoretical width of a 95% confidence interval for W, computed using (5). The quantity $\bar{c}(n, 0.95)$ is the estimate of the true coverage, i.e., the fraction of the 100 estimated 95% confidence intervals which contain W. The following observations can be made based on the results in Table 2:

1. Even for small values of n (e.g., $n = 50$), the bias of $W(n)$ is not significant. (The bias is estimated by $\overline{W}(n) - W$.)
2. The width of the estimated 95% confidence interval for W is less than the approximate theoretical width and the estimated width approaches the approximate theoretical width as n increases. Recall that the approximate theoretical width is equal to the estimated width with $V(n)$ replaced by V and $\nu(n)$ replaced by $\mathrm{E}[\nu]$.
3. The estimated coverage is below 0.95 for small n, but the agreement is better for larger values of n. For $n =$

500 reasonably good agreement is obtained for $\rho = 0.2$ and $\rho = 0.5$. For $\rho = 0.8$ the estimated coverage is still quite low for $n = 1000$.

To summarize these observations, it appears that for low and medium traffic intensities confidence intervals estimated on the basis of the regenerative stochastic structure of the M/G/1 queue are reasonably valid (i.e., the estimated coverage does not differ greatly from 0.95) for $n \geq 500$. For high traffic intensities a much larger number of tours must be simulated in order to obtain reasonably valid confidence intervals. Observing in Table 1 that several thousand tours must be simulated at high traffic intensities if a narrow confidence interval (width equal to $0.1W$) is required, one might expect that if (at a high traffic intensity) the number of tours is sufficiently large to yield a narrow confidence interval for W, then the estimated confidence interval will be reasonably valid. (The M/G/1 queue was not simulated for several thousand tours at $\rho = 0.8$ due to the large amount of computer time required for 100 replications. It took almost 20 minutes of computer time on a large computer to run 100 replications for $n = 1000$ and $\rho = 0.8$.)

Comments

The method for estimating confidence intervals, reviewed in this paper, can only be applied if certain requirements are met: 1) the stochastic system being simulated must be regenerative and 2) certain random variables associated with a tour must have finite first two moments. It is known that these requirements are met for the M/G/1 queue if and only if the traffic intensity is less than one and the service time has finite first four moments. The method, however, is of little interest for analytically tractable queuing systems (such as the M/G/1 queue) except, perhaps, for systems whose analytic solution requires excessive computation.

In general, one must be concerned with whether these requirements are met. It has been proven that certain queuing systems are regenerative even though response variables for the system cannot be obtained analytically (e.g., see [6, 11] in which open systems are considered and [12] in which closed systems are considered; it has been proven for the closed systems in [12] that requirement (2) is met also). For other queuing systems it may not be possible to prove that the system is regenerative, even though intuitively it appears to be. For example, in a subsequent paper [5] a complex open queuing system is considered with Poisson arrivals and independent service times. It is easy to show that the system in [5] stochastically restarts whenever a customer arrives at the empty system and, thus, the time of arrival of a customer at the empty system is a regeneration point. (This property holds for many open queuing systems.) While it is plausible for this system that at sufficiently low input rates the time between successive regeneration points is finite with probability one (i.e., the system is regenerative), no proof could be found. Nevertheless, requirements (1) and (2) were conjectured to hold and the regenerative method was successfully applied to this system.

References

1. J. P. C. Kleijnen, *Statistical Techniques in Simulation*, Part II, Marcel Dekker, Inc., New York, 1975.
2. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems: III. Regenerative Processes and Discrete-Event Simulations," *Oper. Res.* **23,** 33 (1975).
3. D. L. Iglehart, "Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators," *Naval Res. Logist. Quart.*, to be published.
4. S. S. Lavenberg and G. S. Shedler, "Derivation of Confidence Intervals for Work-Rate Estimators in a Closed Queueing Network," *SIAM J. Comput.* **4**, 108 (1975).
5. S. S. Lavenberg and D. R. Slutz, "Regenerative Simulation of a Queuing Model of an Automated Tape Library," *IBM J. Res. Develop.* **19,** 463 (1975; this issue).
6. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems, I: General Multiserver Queues," *J. ACM* **21,** 103 (1974).
7. G. S. Fishman, "Estimation in Multiserver Queuing Simulations," *Oper. Res.* **22,** 72 (1974).
8. A. M. Law, "Efficient Estimators for Simulated Queuing Systems," *ORC 74-7*, Operations Research Center, University of California, Berkeley, 1974.
9. W. Feller, *An Introduction to Probability Theory and its Applications*, Volume I, John Wiley & Sons, Inc., New York, 1968.
10. K. L. Chung, *A Course in Probability Theory*, Harcourt, Brace & World, Inc., New York, 1968.
11. S. S. Lavenberg, "Stability and Maximum Departure Rate of Certain Open Queueing Networks Having Finite Capacity Constraints," (to appear *IBM Research Report*, 1975).
12. S. S. Lavenberg, "Efficient Estimation via Simulation of Work-Rates in Closed Queueing Networks," *Proceedings in Computational Statistics*, Physica Verlag, Vienna, 1974, pp. 353-362.

Received February 4, 1975; revised April 15, 1975

The authors are located at the IBM Research Division Laboratory, Monterey and Cottle Roads, San Jose, California 95193.