S. S. Lavenberg
C. H. Sauer

Sequential Stopping Rules for the Regenerative Method of Simulation

Abstract: We consider the estimation via simulation of confidence intervals for steady-state response variables for stochastic systems which have a regenerative stochastic structure. Sequential stopping rules are investigated which allow the ratio of the width to the midpoint of an estimated confidence interval to be specified ahead of time. We prove that the resulting confidence intervals are valid asymptotically as the relative width decreases to zero. For various relative widths we empirically investigate the validity of the confidence intervals obtained when the stopping rules are applied to the simulation of queuing systems having a regenerative stochastic structure. For the queuing systems and response variables considered, a relative width of 0.05 is found to be sufficiently small to yield valid confidence intervals in almost all cases. In addition, we empirically compare the sequential stopping rules with a fixed stopping rule.

Introduction

When simulating a stochastic system such as a queuing system, in order to estimate steady-state response variables, it is desirable to obtain both point and confidence interval estimates for the response variables. A distributional theory for estimating confidence intervals is usually based on asymptotic results, so that it is necessary to run the simulation long enough to obtain valid confidence intervals. In addition, it is desirable that the widths of the estimated confidence intervals be sufficiently small that useful conclusions can be drawn from the simulation experiment. The width of an estimated confidence interval can be controlled by the use of an appropriate sequential stopping rule.

In this paper we are concerned with sequential stopping rules for estimating steady-state response variables for regenerative stochastic processes when using the regenerative method of simulation. Papers dealing with the regenerative method of simulation which provide a useful background for this paper are cited in [1-6]. Since the regenerative method of simulation involves estimating the ratio of the mean values of two dependent random variables from a sequence of independent pairs of observations of the random variables, we will be concerned with sequential stopping rules for such ratio estimation. Previously, Chow and Robbins [7] considered a sequential stopping rule for estimating the mean value of a single random variable from a sequence of independent observations of the random variable. This rule results in a confidence interval of specified width. They derived asymptotic properties of the stopping rule as the width decreases to zero, including the asymptotic validity of the resulting confidence interval.

In the next section of this paper, we briefly describe the regenerative method of simulation. Sequential stopping rules, which control the relative width of an estimated confidence interval (i.e., the ratio of the width of the interval to its midpoint) and can be used in conjunction with regenerative simulation, are proposed in the third section, where we also derive the asymptotic validity of the resulting confidence intervals as the relative width decreases to zero. The fourth section contains empirical results on the finite sample properties of the stopping rules applied to the regenerative simulation of queuing systems. For the queuing systems and response variables considered, we investigate how small the relative width must be in order to obtain valid confidence intervals. In addition, we empirically compare the sequential stopping rules with a fixed stopping rule. Our conclusions are presented in the last section.

In a recent paper Robinson [8] independently considered the application of sequential stopping rules to regenerative simulation and addressed some of the questions we consider in this paper. However, the asymptotic validity of the resulting confidence intervals was not properly established in that paper. (In particular, the limit theorem due to Anscombe [9], which was used in [8] for establishing the asymptotic validity, was proved under certain restrictive assumptions which were not

demonstrated to hold for the sequential stopping rules.) In addition, our paper includes much more extensive empirical studies for queuing systems than did [8].

Regenerative simulation

Let $\{\mathbf{V}(t):t \geq 0\}$ be a vector-valued continuous parameter stochastic process that assumes values in k-dimensional Euclidean space. We assume that $\{\mathbf{V}(t):t \geq 0\}$ is a *regenerative* process with an infinite sequence of regeneration points $\{t_i:i=1, 2, \cdots\}$, where $0 \leq t_1 < t_2 < \cdots$. Informally, this means that $\{\mathbf{V}(t):t_i \leq t < t_{i+1}\}$, the evolution in time of the process between two successive regeneration points, is a statistically independent probabilistic replica of the evolution in time of the process between any two other successive regeneration points. The function $\{\mathbf{V}(t):t_i \leq t < t_{i+1}\}$ is called the ith *tour*, $i = 1, 2, \cdots$. Let $X_i = t_{i+1} - t_i$ denote the duration of the ith tour. The random variables $\{X_i:i=1, 2, \cdots\}$ are iid (independent and identically distributed).

Under certain mild regularity conditions [3] it can be shown that if X_1 is not a discrete random variable and if $\mathrm{E}[X_1] < \infty$, then

$$\lim_{t\to\infty} P(\mathbf{V}(t) \leq \mathbf{v}) = P(\mathbf{V} \leq \mathbf{v}),$$

i.e., $\{\mathbf{V}(t):t \geq 0\}$ has a limiting probability distribution. The random vector $\mathbf{V}$ is the so-called steady-state vector. Let f be a real-valued non-negative measurable function defined on k-dimensional Euclidean space, and let

$$Y_i = \int_{t_i}^{t_{i+1}} f[\mathbf{V}(t)]\, dt.$$

The random variables $\{Y_i:i=1, 2, \cdots\}$ are iid. If X_1 is not a discrete random variable, if $\mathrm{E}[X_1] < \infty$ and if $\mathrm{E}[f(\mathbf{V})] < \infty$, then, under certain mild regularity conditions [3], it can be shown that

$$\mathrm{E}[f(\mathbf{V})] = \lim_{t\to\infty} \frac{1}{t}\int_0^t f[\mathbf{V}(u)]\, du = \mathrm{E}[Y_1]/\mathrm{E}[X_1], \quad P1 \text{ (with probability one)}. \tag{1}$$

Furthermore, let g be a real-valued non-negative measurable function defined on $2k$-dimensional Euclidean space, where $g(\mathbf{x}, \mathbf{y})$ is a cost incurred when $\mathbf{V}(t)$ undergoes a transition from state $\mathbf{x}$ to state $\mathbf{y}$. Let $C(t, g)$ be the sum of the costs incurred for all transitions which occur in the time interval $[0, t)$, and let

$$C_i(g) = C(t_{i+1}, g) - C(t_i, g),$$

i.e., $C_i(g)$ is the total cost incurred during the ith tour. The random variables $\{C_i(g):i = 1, 2, \cdots\}$ are iid. If $\mathrm{E}[X_1] < \infty$ and $\mathrm{E}[C_i(g)] < \infty$, it can be shown that [3]

$$\lim_{t\to\infty} \frac{C(t, g)}{t} = \frac{\mathrm{E}[C_1(g)]}{\mathrm{E}[X_1]}, \quad P1. \tag{2}$$

The limit is the average cost per unit time. (The cost considered in [3] is somewhat more general than the cost we consider here.)

Similar results hold for a discrete parameter regenerative stochastic process $\{\mathbf{V}_n:n = 1, 2, \cdots\}$ with an infinite sequence of regeneration points $\{n_i:i = 1, 2, \cdots\}$, where $0 \leq n_1 < n_2 < \cdots$. If $\mathrm{E}[n_2 - n_1] < \infty$ and the probability distribution of $n_2 - n_1$ does not assign all its weight to values that are integer multiples of some integer $I > 1$, then

$$\lim_{n\to\infty} P(\mathbf{V}_n \leq \mathbf{v}) = P(\mathbf{V} \leq \mathbf{v}).$$

Further, if $\mathrm{E}[f(\mathbf{V})] < \infty$, then

$$\mathrm{E}[f(\mathbf{V})] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{V}_i) \qquad P1$$

$$= \mathrm{E}\left[\sum_{i=n_1}^{n_2-1} f(\mathbf{V}_i)\right] / \mathrm{E}[n_2 - n_1].$$

We omit the average cost results. For the sake of simplicity we proceed with our discussion in terms of estimating $\mathrm{E}[f(\mathbf{V})]$ for a continuous parameter regenerative stochastic process.

In order to estimate the steady-state response variable $r = \mathrm{E}[f(\mathbf{V})]$ by the regenerative method of simulation, we simulate the regenerative process $\{\mathbf{V}(t):t \geq 0\}$ and collect the sequence of pairs of observations $\{(X_i, Y_i):i = 1, 2, \cdots\}$. The pair (X_i, Y_i) is defined solely with respect to the ith tour and $\{(X_i, Y_i):i = 1, 2, \cdots\}$ is a sequence of iid pairs of non-negative random variables. Point and confidence interval estimates for r are obtained from these observations as follows: Let

$$X(n) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$Y(n) = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$r(n) = \frac{Y(n)}{X(n)},$$

$$S_x^2(n) = \frac{1}{n-1}\sum_{i=1}^{n} [X_i - X(n)]^2,$$

$$S_y^2(n) = \frac{1}{n-1}\sum_{i=1}^{n} [Y_i - Y(n)]^2,$$

$$S_{xy}(n) = \frac{1}{n-1}\sum_{i=1}^{n} [X_i - X(n)][Y_i - Y(n)], \text{ and}$$

$$S^2(n) = S_y^2(n) - 2r(n)S_{xy}(n) + r^2(n)S_x^2(n).$$

Then it can be shown that

$$\lim_{n\to\infty} r(n) = r, \qquad P1, \tag{3}$$

and, if $E[X_1^2] < \infty$ and $E[Y_1^2] < \infty$, then

$$\lim_{n\to\infty} P(n^{\frac{1}{2}}X(n)[r(n) - r]/S(n) \leq t) = \phi(t), \tag{4}$$

where

$$\phi(t) = \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \int_{-\infty}^{t} \exp(-u^2/2)\, du.$$

From (3), the point estimator $r(n)$ is a strongly consistent estimator of r. From (4) it follows that

$$\lim_{n\to\infty} P(r(n) - \delta(n, \alpha) < r < r(n) + \delta(n, \alpha)) = \alpha,$$

where

$$\delta(n, \alpha) = \phi^{-1}[(1 + \alpha)/2]S(n)/n^{\frac{1}{2}}X(n). \tag{5}$$

Thus, if n tours are simulated and n is sufficiently large, the interval

$$(r(n) - \delta(n, \alpha), r(n) + \delta(n, \alpha))$$

is an approximately $100 \times \alpha$ percent confidence interval for r. Other point and confidence interval estimators which have been considered in conjunction with the regenerative method are discussed in [4].

Specifying n ahead of time, i.e., using a fixed stopping rule for the simulation, has the disadvantage that the width $2\delta(n, \alpha)$ or the relative width $2\delta(n, \alpha)/r(n)$ of the estimated confidence interval cannot be specified in advance. In the remainder of this paper we derive theoretical results for and empirically investigate sequential stopping rules for the regenerative method of simulation. These rules allow the relative width of the estimated confidence interval to be specified ahead of time.

Sequential stopping rules

We next define a sequential stopping rule that terminates a regenerative simulation when the relative width of the estimated confidence interval falls below a specified value. Recall that $\{(X_i, Y_i): i = 1, 2, \cdots\}$ are iid pairs of non-negative random variables where (X_i, Y_i) is observed on the ith tour. The response variable to be estimated is $r = E[Y_1]/E[X_1]$. We assume that $0 < E[X_1^2] < \infty$, $0 < E[Y_1^2] < \infty$ and $P(Y_1 = rX_1) < 1$. The last assumption excludes the trivial case $r = Y_1/X_1$ with probability one, for which the exact value of r can be obtained by observing a single tour.

For any $\gamma > 0$, and $0 < \alpha < 1$, let

$$N(\gamma, \alpha) = \min\{n: n \geq 2;$$
$$S(n) > 0;\ 2\delta(n, \alpha)/r(n) < \gamma\}. \tag{6}$$

We stop the simulation when $N(\gamma, \alpha)$ tours have been simulated. This sequential stopping rule is not new and was mentioned by Iglehart [5]. A generalization of this stopping rule, which we discuss later in this section, was employed in the queuing network simulator APLOMB [10]. Similar rules have been independently investigated by Robinson [8] (see our comments at the end of the first section). We will establish that for γ sufficiently small, the interval

$$I(\gamma, \alpha) = (r(N(\gamma, \alpha)) - \delta(N(\gamma, \alpha), \alpha),$$
$$r(N(\gamma, \alpha)) + \delta(N(\gamma, \alpha), \alpha)), \tag{7}$$

where $\delta(n, \alpha)$, as given by (5), is approximately a $100 \times \alpha$ percent confidence interval for r. The relative width of this interval is less than γ.

Let (Ω, A, P) denote the underlying probability space for the regenerative process as defined in [3]. Then $N(\gamma, \alpha)$ is a random variable defined on this probability space. For each $\omega \in \Omega$, $N(\gamma, \alpha, \omega)$ is nondecreasing as γ decreases. In Appendix 1, we prove the following lemmas.

Lemma 1

$$\lim_{\gamma\to 0+} N(\gamma, \alpha) = \infty, \qquad P1.$$

Lemma 2

$$\lim_{\gamma\to 0+} X(N(\gamma, \alpha)) = E[X_1], \qquad P1,$$

$$\lim_{\gamma\to 0+} Y(N(\gamma, \alpha)) = E[Y_1], \qquad P1,$$

$$\lim_{\gamma\to 0+} r(N(\gamma, \alpha)) = r, \qquad P1, \text{ and}$$

$$\lim_{\gamma\to 0+} S(N(\gamma, \alpha)) = S, \qquad P1,$$

where $S^2 = \text{var}(Y_1 - rX_1)$.

Lemma 3

$$\lim_{\gamma\to 0+} \gamma^2 N(\gamma, \alpha) = D, \qquad P1,$$

where $D = \{2\phi^{-1}[(1 + \alpha)/2]S/E[Y_1]\}^2$.

Note that since $P(Y_1 - rX_1 = 0) < 1$, it follows that $S > 0$ and, hence, $D > 0$. Lemma 3 is the key result which we use to prove the following theorem.

Theorem 1

$$\lim_{\gamma\to 0+} P(N(\gamma, \alpha)^{\frac{1}{2}}X(N(\gamma, \alpha))$$
$$\times [r(N(\gamma, \alpha)) - r]/S(N(\gamma, \alpha)) \leq t) = \phi(t).$$

Proof Let $Z_i = Y_i - rX_i$. Then $\{Z_i: i = 1, 2, \cdots\}$ is a sequence of iid random variables, each with mean zero and variance S^2. Let

$$Z(n) = \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

Using Lemma 3 we can proceed in a manner similar to that used by Chung [11], pp. 197–199, in proving a

central limit theorem in the case of a random number of terms, and establish that

$$\lim_{\gamma \to 0+} P(N(\gamma, \alpha)^{\frac{1}{2}} Z(N(\gamma, \alpha))/S \leq t) = \phi(t).$$

Since $Z(N(\gamma, \alpha)) = X(N(\gamma, \alpha))[r(N(\gamma, \alpha)) - r]$, and

$$\lim_{\gamma \to 0+} S(N(\gamma, \alpha)) = S, \qquad P1,$$

the application of Theorem 4.4.8 in Chung [11] completes the proof. □

It follows from Theorem 1 that for γ sufficiently small, the interval $I(\gamma, \alpha)$ given by (7) is approximately a $100 \times \alpha$ percent confidence interval for r. In order to implement the sequential stopping rule it is necessary to recompute the relative width of the confidence interval after each additional tour is simulated. However, we can modify the stopping rule so that this computation is done every K tours, where $K > 1$, instead of every tour. Furthermore, it may be desirable to guarantee that at least a fixed minimum number of tours is always simulated. For any positive integers $K \geq 1$ and $L \geq 2$, and any $\gamma > 0$, let

$$N(K, L, \gamma, \alpha) = \min\{Kn: Kn \geq L;\ S(Kn) > 0;\ 2\delta(Kn, \alpha)/r(Kn) < \gamma\}. \qquad (8)$$

We stop the simulation when $N(K, L, \gamma, \alpha)$ tours have been simulated; $N(K, L, \gamma, \alpha)$ is never less than L. Furthermore, the relative width is only computed every K tours. Lemmas 1–3 and Theorem 1 can be shown to hold for this sequential stopping rule. Thus, for γ sufficiently small,

$$(r(N(K, L, \gamma, \alpha)) - \delta(N(K, L, \gamma, \alpha), \alpha),\ r(N(K, L, \gamma, \alpha)) + \delta(N(K, L, \gamma, \alpha), \alpha))$$

is approximately a $100 \times \alpha$ percent confidence interval for r.

It may also be desirable to place a fixed upper bound on the number of tours that are simulated, thus combining aspects of a fixed and sequential stopping rule. For any positive integer $M > L$, let

$$N(K, L, M, \gamma, \alpha) = \min(N(K, L, \gamma, \alpha), M).$$

It can be proved from previous results for the pure fixed stopping rule and for the pure sequential stopping rule that for γ sufficiently small and M sufficiently large,

$$(r(N(K, L, M, \gamma, \alpha)) - \delta(N(K, L, M, \gamma, \alpha), \alpha),\ r(N(K, L, M, \gamma, \alpha)) + \delta(N(K, L, M, \gamma, \alpha), \alpha))$$

is approximately a $100 \times \alpha$ percent confidence interval for r. This last stopping rule was implemented in APLOMB [10].

In the sequential stopping rules we have discussed, we do not stop after n tours if $S(n) = 0$, and hence, $\delta(n, \alpha)/r(n) = 0$. It is straightforward to show that $S(n) = 0$ if and only if $Y_1/X_1 = Y_2/X_2 = \cdots = Y_n/X_n$. Therefore, for any $n \geq 2$, $S(n) > 0$ if $S(2) > 0$, and $P(S(n) > 0, n = 2, 3, \cdots) = P(S(2) > 0) = P(Y_1/X_1 \neq X_2/Y_2)$. If $P(Y_1/X_1 = Y_2/X_2) > 0$, then $P(S(2) = 0) > 0$. If, for example, we replace (6) with

$$N'(\gamma, \alpha) = \min\{n: n \geq 2;\ 2\delta(n, \alpha)/r(n) < \gamma\},$$

where the restriction $S(n) > 0$ is omitted, then if $p^* = P(Y_1/X_1 = Y_2/X_2) > 0$, it follows that $P(N'(\gamma, \alpha) = 2) \geq p^* > 0$, where p^* does not depend on γ. Hence,

$$P(\lim_{\gamma \to 0+} N'(\gamma, \alpha) = \infty) \leq 1 - p^*,$$

i.e., Lemma 1 does not hold. We have avoided this situation by stopping after n tours only if $S(n) > 0$. If $P(Y_1/X_1 = Y_2/X_2) = 0$, which holds, for example, if Y_1/X_1 is absolutely continuous, then we do not need this restriction. Another way to avoid this problem is to replace (6) with

$$N''(\gamma, \alpha) = \min\{n: n \geq 2;\ [2\delta(n, \alpha)/r(n)] + 1/n < \gamma\}.$$

Empirical studies

In this section we will present empirical results on the actual coverage obtained when using the sequential rules in the regenerative simulation of queuing systems that have known analytic solutions. We also empirically compare the sequential rules with a fixed stopping rule. We describe the queuing systems, the response variables to be considered, and how the regenerative method can be applied to estimate these response variables.

- *M/G/1 queue*

We consider an M/G/1 queue with Poisson arrivals at rate λ and general service times having mean $1/\mu$, finite fourth moment, and probability distribution function $F(t)$. Let Q_n denote the queuing time (waiting time plus service time) for the nth customer to arrive. If $\rho = \lambda/\mu < 1$, then $\{Q_n: n = 1, 2, \cdots\}$ is a discrete parameter regenerative stochastic process with regeneration points $\{n_i: i = 1, 2, \cdots\}$, where n_i is the serial number of the ith customer which arrives to find the system empty [1]. By letting Q denote the mean steady-state queuing time, the regenerative method can be applied to obtain point and confidence interval estimates for the response variable

$$Q = \mathrm{E}\left[\sum_{i=n_1}^{n_2-1} Q_i\right] / \mathrm{E}[n_2 - n_1].$$

- *Cyclic queue with hyperexponential service times*

We consider a cyclic queuing system that consists of two service centers, each of which is a single server queue

with customers served in the order of their arrival (Fig. 1). The number of customers in the system is fixed and equal to N. All service times are mutually independent random variables, and service times at service center i are iid with mean $1/\mu_i$ and probability distribution function

$$F_i(t) = p_i[1 - \exp(-2p_i\mu_i t)] + (1 - p_i) \times \{1 - \exp[-2(1 - p_i)\mu_i t]\}, \qquad t \geq 0, \quad (9)$$

where $p_i = [C_i^2 + 1 - (C_i^4 - 1)^{\frac{1}{2}}]/[2(C_i^2 + 1)]$, and $C_i > 1$ is the coefficient of variation of the service times at service center i. By using the method of stages [12], a service time at service center i can be represented as two independent exponential stages of service in parallel.

For any $t \geq 0$, let $\mathbf{V}(t) = (n_1(t), s_1(t), n_2(t), s_2(t))$, where $n_i(t)$ is the number of customers in service center i at time t, $s_i(t)$ is the stage of service of the customer in service at service center i at time t if $n_i(t) > 0$, and $s_i(t) = 0$ if $n_i(t) = 0$. Then $\{\mathbf{V}(t) : t \geq 0\}$ is a continuous parameter finite-state irreducible Markov process. Hence, $\{\mathbf{V}(t) : t \geq 0\}$ is a regenerative process with regeneration points $\{t_i : i = 1, 2, \cdots\}$, which are the successive times at which the process enters a fixed state $\mathbf{v}^*$, and $\mathrm{E}[t_2 - t_1] < \infty$ [2]. We call $\mathbf{v}^*$ the *tour-defining state*. By choosing appropriate functions f, the regenerative method can be used to estimate such response variables as the mean steady-state number of customers in service center 1. Here, however, we will be concerned with estimating a different quantity.

We call the time between successive arrivals by a customer at service center 1 a *cycle time*. Let T_i denote the cycle time for the ith customer to arrive at service center 1 (i.e., the ith customer to complete service at service center 2). We wish to estimate the average cycle time

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} T_i,$$

provided this limit exists. The sequence $\{T_i : i = 1, 2, \cdots\}$ is a discrete parameter stochastic process, but we have *not* been able to find an infinite sequence of regeneration points for this process. However, we show in Appendix 2 that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} T_i = T < \infty, \qquad P1,$$

and by Little's formula [13], that

$$T = N/\Lambda,$$

where

$$\Lambda = \lim_{t\to\infty} \frac{A(t)}{t}, \qquad P1, \quad (10)$$

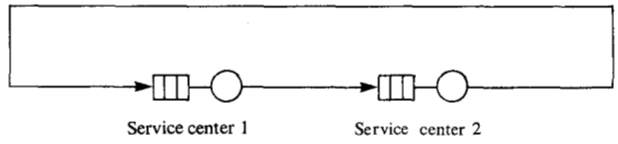

Figure 1 Cyclic queuing system.

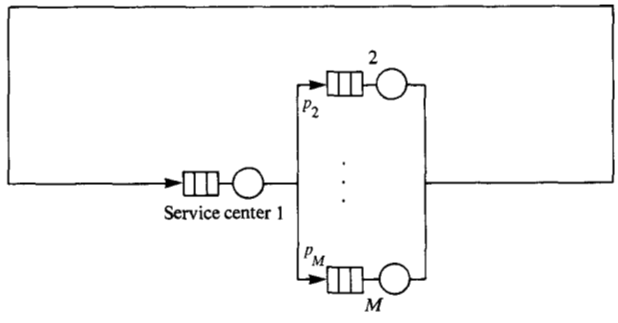

Figure 2 Central server model.

and $A(t)$ is the number of arrivals at service center 1 in the time interval $[0, t)$. The term $A(t)$ counts all transitions from states with k customers in service center 1 to states with $k + 1$ customers in service center 1, for $k = 0, 1, \cdots, N - 1$. Since $\{\mathbf{V}(t) : t \geq 0\}$ is a finite state irreducible Markov process, $\mathrm{E}[A(t_2) - A(t_1)] < \infty$. It follows from (2) that

$$\Lambda = \frac{\mathrm{E}[A_1]}{\mathrm{E}[X_1]},$$

where $X_i = t_{i+1} - t_i$, and $A_i = A(t_{i+1}) - A(t_i)$. Hence,

$$T = \frac{N\mathrm{E}[X_1]}{\mathrm{E}[A_1]},$$

and the regenerative method can be applied to obtain point and confidence interval estimates for the average cycle time.

• *Central server model with exponential service times* We consider the closed queuing system shown in Fig. 2, consisting of M service centers, each of which is a single server queue with customers served in the order of their arrival. The number of customers is fixed and equal to N. A customer completing service at service center 1 immediately enters service center i with branching probability $p_i > 0$, $i = 2, \cdots, M$, where

$$\sum_{i=2}^{M} p_i = 1;$$

a customer completing service at service center i, $i = 2, \cdots, M$, immediately enters service center 1. All service

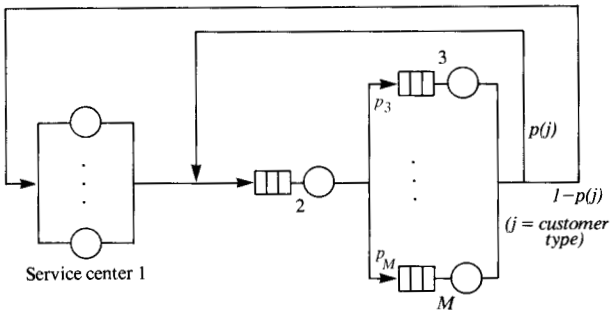

Figure 3 Closed queuing system with two types of customers.

times are mutually independent random variables and service times at service center i are iid and are exponentially distributed with mean $1/\mu_i$. (This queuing system was called the central server model by Buzen [14], who proposed it as a simple model of a multiprogrammed computer system having a fixed level N of multiprogramming, a single processor, and $M-1$ input-output devices. Service center 1, called the central server, represents the processor.)

For any $t \geq 0$, let $\mathbf{V}(t) = (n_1(t), \cdots, n_M(t))$, where $n_i(t)$ is the number of customers in service center i at time t. Then $\{\mathbf{V}(t) : t \geq 0\}$ is a continuous parameter finite state irreducible Markov process, and hence is a regenerative process. The regeneration points $\{t_i : i = 1, 2, \cdots\}$ are the successive times at which the process enters a fixed state $\mathbf{v}^*$, the tour-defining state, and $\mathrm{E}[t_2 - t_1] < \infty$. We call the time between successive arrivals by a customer at service center 1 a cycle time, and let T_i denote the cycle time for the ith customer to arrive at service center 1. Then, as for the cyclic queue, it can be shown that the average cycle time exists and is finite, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} T_1 = T < \infty, \qquad P1,$$

and that

$$T = \frac{N\mathrm{E}[X_1]}{\mathrm{E}[A_1]},$$

where $X_i = t_{i+1} - t_i$, $A_i = A(t_{i+1}) - A(t_i)$, and $A(t)$ is the number of customers arriving at service center 1 in the time interval $[0, t)$. The regenerative method can be applied to obtain point and confidence interval estimates for the average cycle time.

•. *Closed queuing system with two types of customers* We consider the closed queuing system consisting of M service centers (Fig. 3). There are two types of customers in the network. The number of type 1 customers is fixed and equal to N_1 and the number of type 2 customers is fixed and equal to N_2. Service center 1 has $N = N_1 + N_2$ identical parallel servers; hence a customer never has to wait for a server at service center 1 to become free. Service center 2 is a single server processor sharing queue [15]; i.e., all customers present at service center 2 receive service simultaneously, and, if there are n customers present, each customer is served at $(1/n)$th of the server's rate. Service centers $3, \cdots, M$ are single server queues with customers served in the order of their arrival. A customer completing service at service center 1 immediately enters service center 2; a customer of either type completing service at service center 2 immediately enters service center i with probability $p_i > 0$, $i = 3, \cdots, M$, where

$$\sum_{i=3}^{M} p_i = 1;$$

a type j customer completing service at service center i, $i = 3, \cdots, M$, immediately enters service center 2 with probability $p(j) > 0$, or immediately enters service center 1 with probability $1 - p(j) > 0$. All service times are mutually independent random variables. For $i = 1, 2$, service times at service center i for type j customers are iid and exponentially distributed with mean $1/(\mu_{ij})$; for $i = 3, \cdots, M$, service times at service center i are iid and are exponentially distributed with mean $1/\mu_i$ for both types of customers. This queuing system is illustrative of fairly complex queuing models of interactive computer systems. Service center 1 represents a collection of currently active terminals; service center 2 represents a processor, which is scheduled in a round-robin manner; service centers $3, \cdots, M$ represent input-output devices, each of which is scheduled on a first-come, first-served basis. There are two types of users at the terminals and the two types differ in their think times (service times at service center 1), processor times between input-output requests (service times at service center 2), and number of input-output requests per interaction (number of visits to service centers $3, \cdots, M$ between visits to service center 1). The particular assumptions we have made about this model allow us to compute the average response time defined later [16].

For any $t \geq 0$, let $\mathbf{V}(t) = (\mathbf{V}_1(t), \cdots, \mathbf{V}_M(t))$, where $\mathbf{V}_i(t)$ is a list of the types of the customers in service center i at time t in the order of their arrival at the service center. It can be shown that $\{\mathbf{V}(t) : t \geq 0\}$ is a continuous parameter finite-state irreducible Markov process, and, hence, is a regenerative process. The regeneration points $\{t_i : i = 1, 2, \cdots\}$ are the successive times at which the process enters a fixed state $\mathbf{v}^*$, the tour-defining state, and $\mathrm{E}[t_2 - t_1] < \infty$. We call the time from when a customer leaves service center 1 until that customer next returns to service center 1 a *response time*. Let R_i denote the response time for the ith customer to leave service

center 1. Then it can be shown that the average response time exists and is finite, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R_i = R < \infty, \qquad P1,$$

and that

$$R = \frac{\mathrm{E}[L_1]}{\mathrm{E}[A_1]},$$

where $A_i = A(t_{i+1}) - A(t_i)$, $L_i = \int_{t_i}^{t_{i+1}} n(t)\ dt$, $A(t)$ is the number of customers that leave service center 1 in the time interval $[0, t)$, and $n(t)$ is the total number of customers in service centers $2, \cdots, M$ at time t. Thus, the regenerative method can be used to obtain point and confidence interval estimates for the average response time.

The empirical studies we next describe were carried out using APLOMB [10], a FORTRAN program which simulates a broad class of queuing systems having a regenerative stochastic structure and uses the regenerative method to obtain point and confidence intervals for steady-state response variables. Integer random numbers are generated in APLOMB by the multiplicative congruential generator $W_n = 7^5\ W_{n-1} (\mathrm{mod}\ 2^{31} - 1)$ discussed in [17]. A realization of an exponentially distributed random variable having mean $1/\mu$ is obtained by the transformation $-(1/\mu)\ \ln[W_n/(2^{31} - 1)]$. A realization of an Erlang-k random variable is obtained by summing k realizations of an exponential random variable. A hyperexponential random variable having the probability distribution function given in (9) can be represented by the method of stages as shown in Fig. 4, where U_1 and U_2 are independent exponentially distributed random variables having means $1/2(1 - p_i)\mu_i$ and $1/2p_i\mu_i$ respectively, and $\delta = (1 - 2p_i)p_i/(1 - p_i)$. In APLOMB a realization of a hyperexponential random variable is obtained by using this representation from a realization of U_1, a realization of a binary-valued random variable and, if necessary, a realization of U_2.

We conducted two sets of experiments to study the finite sample properties of sequential stopping rules, and a third set of experiments to compare the finite sample properties of fixed and sequential stopping rules. Each experiment consisted of I independent replications of a regenerative simulation, where $I = 100$ unless we state otherwise. Each simulation was started in the tour-defining state. For each replication we determined whether or not the known response variable r was contained in the estimated confidence interval for r. From the I replications comprising an experiment we obtained point and 90-percent confidence interval estimates for the true coverage of the confidence interval for r. (The confidence interval for the true coverage was estimated as described in Appendix 3.) When the confidence interval for the true coverage contains the desired coverage α, we say that the confidence interval for r is valid. The sequential stopping rule we used is to simulate $N(K, L, \gamma, \alpha)$ tours, where $N(K, L, \gamma, \alpha)$ is given by (8), $L = 10$ and $\alpha = 0.9$. In addition, for each experiment we obtained point and 90-percent confidence interval estimates for the relative bias of the point estimate for r. [If r' is the point estimate of r, the relative bias in percent is equal to $100(E[r'] - r)/r$]. The confidence interval estimate was

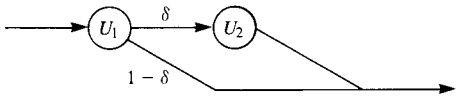

Figure 4 Representation of hyperexponential random variable.

Table 1a M/G/1 queue: description of systems.

System	λ	μ	C	ρ	Q
1	0.500	1.000	0.5	0.50	1.625
2	0.500	1.000	1.0	0.50	2.000
3	0.500	1.000	2.0	0.50	3.500

Table 1b M/G/1 queue: empirical results.

System	K	γ	*Coverage*	*Number of Tours*	*Relative Bias (%)*
1	10	0.10	0.88 (0.82, 0.92)	1414 (1337, 1491)	−0.9 (−1.4, −0.4)
1	50	0.10	0.86 (0.79, 0.91)	1442 (1373, 1511)	−0.8 (−1.3, −0.3)
1	100	0.10	0.90 (0.84, 0.94)	1472 (1394, 1550)	−0.8 (−1.4, −0.3)
2	10	0.10	0.85 (0.78, 0.90)	4455 (4285, 4625)	−0.7 (−1.3, −0.2)
2	50	0.10	0.90 (0.84, 0.94)	4489 (4320, 4657)	−0.7 (−1.2, −0.2)
2	100	0.10	0.86 (0.79, 0.91)	4566 (4383, 4749)	−0.7 (−1.3, −0.2)
3	10	0.10	0.86 (0.79, 0.91)	31652 (30681, 32624)	−0.8 (−1.4, −0.2)
3	50	0.10	0.84 (0.77, 0.89)	31886 (30909, 32863)	−0.6 (−1.2, −0.1)
3	100	0.10	0.85 (0.78, 0.90)	31503 (30532, 32474)	−0.8 (−1.4, −0.3)

obtained from the I independent samples of $(r' - r)/r$ by using the t-statistic with $I - 1$ degrees of freedom. For each experiment using a sequential stopping rule we obtained point and 90-percent confidence interval estimates for $E[N(K, L, \gamma, \alpha)]$ in the same manner.

The purpose of the first set of experiments was to study the effect of different values of K, and K was allowed to have the values 10, 50 and 100. The fixed value $\gamma = 0.1$ was used. The systems studied were M/G/1 queues with $\lambda = 0.5$, $\mu = 1$ and three different forms for the service time probability distribution function $F(t)$. The three forms are Erlang-k with coefficient of variation equal to 0.5, i.e., $k = 4$, exponential, and the hyperexponential form given in (9) with coefficient of variation equal to 2. The description of these three systems is summarized in Table 1a, where C denotes the coefficient of variation of the service times and Q is the known mean steady-state queuing time. Results of the nine experiments are given in Table 1b. The first column identifies one of the systems in Table 1a. Except for one experiment, valid confidence intervals were obtained for the response variable. Furthermore, we concluded that different values of K have little effect on the results.

The second set of experiments was more extensive and was used to explore the effects of different values of γ, the specified relative width. The value of K used depended on the particular system (see Appendix 4). First we ran experiments for the M/G/1 queue with three different arrival rates and the three forms for the service time distribution discussed above. These nine systems are summarized in Table 2a. Table 2b contains results of 27 experiments with these nine systems using the values 0.3, 0.2, and 0.1 for γ. Valid confidence intervals were obtained for Q in eight of the nine experiments with $\gamma = 0.1$, in four of the nine experiments with $\gamma = 0.2$, and in none of the experiments with $\gamma = 0.3$.

Table 2a M/G/1 queue: description of systems.

System	λ	μ	C	ρ	Q
1	0.250	1.000	0.5	0.25	1.208
2	0.500	1.000	0.5	0.50	1.625
3	0.750	1.000	0.5	0.75	2.877
4	0.250	1.000	1.0	0.25	1.333
5	0.500	1.000	1.0	0.50	2.000
6	0.750	1.000	1.0	0.75	4.003
7	0.250	1.000	2.0	0.25	1.833
8	0.500	1.000	2.0	0.50	3.500
9	0.750	1.000	2.0	0.75	8.507

Table 2b M/G/1 queue: empirical results.

System	K	γ	*Coverage*	*Number of Tours*	*Relative Bias (%)*
1	10	0.30	0.82 (0.75, 0.87)	51 (45, 57)	−1.9 (−4.0, 0.0)
1	10	0.20	0.81 (0.74, 0.87)	128 (118, 138)	−2.1 (−3.2, −0.9)
1	30	0.10	0.91 (0.85, 0.95)	546 (526, 566)	−0.9 (−1.4, −0.5)
2	10	0.30	0.66 (0.58, 0.73)	87 (76, 98)	−7.9 (−10.0, −5.8)
2	20	0.20	0.75 (0.67, 0.81)	278 (254, 302)	−3.3 (−4.4, −2.2)
2	81	0.10	0.88 (0.82, 0.92)	1498 (1415, 1580)	−0.8 (−1.4, −0.3)
3	33	0.30	0.62 (0.54, 0.70)	305 (265, 344)	−8.5 (−10.7, −6.2)
3	74	0.20	0.71 (0.63, 0.78)	1004 (893, 1115)	−4.9 (−6.2, −3.5)
3	297	0.10	0.84 (0.77, 0.89)	5325 (5051, 5599)	−1.0 (−1.5, −0.5)
4	12	0.30	0.84 (0.77, 0.89)	207 (194, 219)	−3.4 (−5.0, −1.8)
4	28	0.20	0.88 (0.82, 0.92)	520 (492, 548)	−1.8 (−2.8, −0.7)
4	112	0.10	0.88 (0.82, 0.92)	2227 (2159, 2294)	−0.9 (−1.4, −0.3)
5	27	0.30	0.75 (0.67, 0.81)	368 (334, 402)	−5.6 (−7.3, −3.9)
5	60	0.20	0.88 (0.82, 0.92)	1083 (1015, 1151)	−1.6 (−2.7, −0.5)
5	243	0.10	0.88 (0.82, 0.92)	4697 (4496, 4898)	−0.5 (−1.0, 0.0)
6	73	0.30	0.74 (0.66, 0.81)	965 (858, 1072)	−5.0 (−7.0, −3.1)
6	165	0.20	0.82 (0.75, 0.87)	2742 (2564, 2921)	−2.2 (−3.3, −1.1)
6	663	0.10	0.86 (0.79, 0.91)	12060 (11606, 12514)	−1.0 (−1.6, −0.5)
7	102	0.30	0.83 (0.76, 0.88)	1799 (1669, 1930)	−3.4 (−5.1, −1.8)
7	229	0.20	0.87 (0.80, 0.92)	4264 (4066, 4462)	−2.2 (−3.3, −1.3)
7	918	0.10	0.88 (0.82, 0.92)	18167 (17664, 18671)	−0.4 (−0.9, 0.0)
8	179	0.30	0.82 (0.75, 0.87)	3047 (2841, 3252)	−3.5 (−5.0, −2.0)
8	404	0.20	0.87 (0.80, 0.92)	7583 (7229, 7937)	−1.3 (−2.3, −0.3)
8	1618	0.10	0.86 (0.79, 0.91)	31826 (30936, 32717)	−0.7 (−1.2, −0.2)
9	319	0.30	0.79 (0.72, 0.85)	4967 (4563, 5370)	−3.5 (−5.3, −1.8)
9	718	0.20	0.83 (0.76, 0.88)	12946 (12065, 13826)	−1.8 (−2.9, −0.7)
9	2872	0.10	0.89 (0.83, 0.93)	56435 (54484, 58385)	−0.9 (−1.4, −0.4)

Table 3a Cyclic queuing systems: description of systems.

System	μ_1	μ_2	C_1	C_2	ρ_1	ρ_2	N	T
1	1.000	1.000	2	2	0.72	0.72	5	6.959
2	1.000	1.000	2	2	0.79	0.79	10	12.585
3	1.000	1.000	4	4	0.65	0.65	5	7.644
4	1.000	1.000	4	4	0.70	0.70	10	14.308
5	1.000	1.000	8	8	0.63	0.63	5	7.902
6	1.000	1.000	8	8	0.66	0.66	10	15.148
7	0.667	1.333	2	2	0.91	0.45	5	8.276
8	0.667	1.333	2	2	0.97	0.48	10	15.543
9	0.667	1.333	4	4	0.83	0.42	5	8.984
10	0.667	1.333	4	4	0.88	0.44	10	17.079
11	0.667	1.333	8	8	0.81	0.40	5	9.304
12	0.667	1.333	8	8	0.82	0.41	10	18.186

Table 3b Cyclic queuing systems: empirical results.

System	K	γ	*Coverage*	*Number of Tours*	*Relative Bias (%)*
1	10	0.30	0.80 (0.73, 0.86)	50 (46, 54)	−4.5 (−5.9, −3.0)
1	10	0.20	0.84 (0.77, 0.89)	125 (119, 131)	−1.9 (−2.9, −0.9)
1	10	0.10	0.91 (0.85, 0.95)	525 (511, 538)	−0.6 (−1.1, −0.1)
2	10	0.30	0.70 (0.62, 0.77)	26 (23, 29)	−4.9 (−6.6, −3.1)
2	10	0.20	0.71 (0.63, 0.78)	58 (53, 64)	−3.7 (−5.4, −2.0)
2	10	0.10	0.89 (0.83, 0.93)	287 (273, 300)	−1.0 (−1.7, −0.3)
3	10	0.30	0.77 (0.69, 0.83)	240 (226, 254)	−6.3 (−9.0, −3.5)
3	10	0.20	0.83 (0.76, 0.88)	592 (572, 612)	−2.0 (−3.6, −0.3)
3	10	0.10	0.91 (0.85, 0.95)	2471 (2442, 2500)	−0.4 (−0.9, 0.0)
4	10	0.30	0.67 (0.59, 0.74)	115 (104, 127)	−4.8 (−7.7, −1.8)
4	10	0.20	0.78 (0.70, 0.84)	312 (292, 331)	−3.8 (−5.8, −1.7)
4	10	0.10	0.91 (0.85, 0.95)	1411 (1386, 1435)	−0.4 (−0.8, 0.0)
5	10	0.30	0.52 (0.44, 0.60)	705 (616, 794)	−24.5 (−29.2, −19.9)
5	10	0.20	0.69 (0.61, 0.76)	2130 (1963, 2296)	−11.5 (−15.4, −7.6)
5	10	0.10	0.87 (0.80, 0.92)	10444 (10244, 10644)	−0.8 (−2.0, 0.2)
6	10	0.30	0.65 (0.57, 0.72)	479 (433, 524)	−16.2 (−20.2, −12.1)
6	10	0.20	0.74 (0.66, 0.81)	1266 (1182, 1350)	−8.0 (−11.4, −4.6)
6	10	0.10	0.85 (0.78, 0.90)	5669 (5520, 5819)	−1.5 (−3.0, 0.0)
7	10	0.30	0.82 (0.75, 0.87)	141 (133, 149)	−3.8 (−5.6, −2.1)
7	10	0.20	0.86 (0.79, 0.91)	336 (325, 348)	−1.5 (−2.6, −0.4)
7	10	0.10	0.93 (0.88, 0.96)	1420 (1399, 1440)	−0.5 (−1.0, −0.1)
8	10	0.30	0.84 (0.77, 0.89)	125 (116, 134)	−3.7 (−5.4, −2.0)
8	10	0.20	0.89 (0.83, 0.93)	308 (295, 321)	−1.4 (−2.4, −0.4)
8	10	0.10	0.89 (0.83, 0.93)	1292 (1270, 1314)	−0.7 (−1.1, −0.2)
9	10	0.30	0.77 (0.69, 0.83)	618 (585, 651)	−5.5 (−8.2, −2.8)
9	10	0.20	0.88 (0.82, 0.92)	1538 (1502, 1573)	−1.1 (−2.2, 0.0)
9	10	0.10	0.91 (0.85, 0.95)	6335 (6264, 6405)	−0.4 (−0.9, 0.0)
10	10	0.30	0.79 (0.72, 0.85)	535 (508, 562)	−4.0 (−6.1, −2.0)
10	10	0.20	0.89 (0.83, 0.93)	1305 (1267, 1342)	−1.6 (−3.0, −0.2)
10	10	0.10	0.89 (0.83, 0.93)	5449 (5383, 5515)	−0.4 (−0.9, 0.0)
11	10	0.30	0.58 (0.50, 0.66)	2093 (1877, 2308)	−18.2 (−22.5, −13.9)
11	10	0.20	0.87 (0.80, 0.92)	6438 (6203, 6674)	−2.6 (−4.5, −0.6)
11	10	0.10	0.90 (0.84, 0.94)	27216 (26933, 27498)	−0.3 (−0.8, 0.1)
12	10	0.30	0.50 (0.42, 0.58)	1710 (1496, 1923)	−23.2 (−27.8, −18.7)
12	10	0.20	0.84 (0.77, 0.89)	5904 (5647, 6160)	−3.7 (−6.0, −1.3)
12	10	0.10	0.85 (0.78, 0.90)	25228 (24965, 25492)	−0.3 (−0.8, 0.1)

Next we ran experiments for the twelve cyclic queuing systems with hyperexponential service times described in Table 3a. The quantity ρ_i in the table is the utilization of service center i, and T is the known average cycle time. We considered both balanced systems (i.e., $\rho_1 = \rho_2$) and unbalanced systems (i.e., $\rho_1 \neq \rho_2$). For each system the values 0.3, 0.2, and 0.1 were used for γ. The tour-defining state for each system was the state for which all customers are in service center 1 and the customer in service is in the first stage of service (i.e., the leftmost stage in Fig. 4). Table 3b contains results of the 36 experiments. Valid confidence intervals for T were ob-

Table 4a Central server model: description of systems.

System	M	μ_1	μ_2	μ_3	p_2	p_3	ρ_1	ρ_2	ρ_3	N	*State*	T
1	3	1.00	0.50	0.50	0.5	0.5	0.67	0.67	0.67	4	1 1 2	6.000
2	3	1.00	0.50	0.50	0.5	0.5	0.80	0.80	0.80	8	1 1 6	10.000
3	3	1.00	0.90	0.10	0.9	0.1	0.67	0.67	0.67	4	1 1 2	6.000
4	3	1.00	0.90	0.10	0.9	0.1	0.80	0.80	0.80	8	1 1 6	10.000
5	3	1.00	0.25	0.25	0.5	0.5	0.38	0.76	0.76	4	1 1 2	10.531
6	3	1.00	0.25	0.25	0.5	0.5	0.44	0.88	0.88	8	1 1 6	18.280
7	3	1.00	0.45	0.05	0.9	0.1	0.38	0.76	0.76	4	1 1 2	10.530
8	3	1.00	0.45	0.05	0.9	0.1	0.44	0.88	0.88	8	5 1 2	18.279
9	3	1.00	1.00	1.00	0.5	0.5	0.91	0.46	0.46	4	3 0 1	4.385
10	3	1.00	1.00	1.00	0.5	0.5	0.99	0.50	0.50	8	7 0 1	8.072
11	3	1.00	1.80	0.20	0.9	0.1	0.91	0.46	0.46	4	3 1 0	4.385
12	3	1.00	1.80	0.20	0.9	0.1	0.99	0.50	0.50	8	7 1 0	8.072

tained in all twelve experiments with $\gamma = 0.1$, in five of the twelve experiments with $\gamma = 0.2$ and in none of the experiments with $\gamma = 0.3$.

Next we ran experiments for the twelve central server models with exponential service times described in Table 4a. Each central server model has three service centers and either four or eight customers. The quantity ρ_i is the utilization of service center i and T is the known average cycle time. Note that we considered various balanced and imbalanced systems with equal and unequal branching probabilities. The three columns labeled *state* give the number of customers in the three service centers in the tour-defining state. The tour-defining state for a system was chosen to minimize the mean number of service completions in the system during a tour. Such a state can be analytically determined for these systems. For each system the values 0.3, 0.2, 0.1 and 0.05 were used for γ. Results of the 48 experiments are given in Table 4b. For nine of the twelve experiments with $\gamma = 0.05$, valid confidence intervals for T were obtained. Valid confidence intervals were obtained in only six of the twelve experiments with $\gamma = 0.1$ and in none of the experiments with $\gamma = 0.2$ or $\gamma = 0.3$.

Finally we ran experiments for the two closed queuing systems with two types of customers described in Table 5a. Each system has six service centers, forty type 1 customers and four type 2 customers. The quantity ρ_i is the utilization of service center i and R is the known average response time. The four columns headed *state* give for the tour-defining state the number of type 1 customers in service center 1, the number of type 2 customers in service center 1, the number of type 1 customers in service center 2, and the number of type 2 customers in service center 2. There are no customers in service centers 3–6 in the tour-defining state. For each system the values 0.2, 0.1, and 0.05 were used for γ. Only fifty replications were run for the experiments in which $\gamma = 0.05$. (For $\gamma = 0.05$ a large amount of computer time was required for each replication.) Results of the six experiments are given in Table 5b. Valid confidence intervals for R were obtained in both of the experiments with $\gamma = 0.05$ and in none of the experiments with $\gamma = 0.1$ or $\gamma = 0.2$.

The third set of experiments used a fixed stopping rule; i.e., each simulation was stopped after a fixed number of tours had been simulated. Experiments were performed with the nine M/G/1 queues described in Table 2a and the second closed system described in Table 5a. The values of the number of tours n used for a system were equal to the point estimates of $\mathrm{E}[N(K, L, \gamma, \alpha)]$ obtained when using the sequential stopping rule for the system for different values of γ. The goal was to compare the fixed and sequential stopping rules for the same mean number of tours. Results of the 27 experiments with the M/G/1 queues are given in Table 6 and results of the three experiments with the closed system are given in Table 7. The confidence intervals obtained by using the fixed stopping rules are valid more frequently than the confidence intervals obtained by using the sequential stopping rules. This is particularly so for the larger relative widths and correspondingly smaller number of tours. The sequential stopping rules have worse small sample behavior than the fixed stopping rules. Of course, with a fixed stopping rule, the relative width of an estimated confidence interval cannot be specified ahead of time.

Conclusions

We conducted extensive empirical studies using sequential stopping rules to estimate confidence intervals having a specified relative width when simulating a variety of queuing systems by the regenerative method. For the systems, response variables and 90-percent level of confidence we considered, a relative width of 0.05 was small enough to yield valid confidence intervals in

Table 4b Central server model: empirical results.

System	*K*	γ	*Coverage*	*Number of Tours*	*Relative Bias (%)*
1	10	0.30	0.78 (0.70, 0.84)	19 (17, 21)	−1.1 (−2.8, 0.4)
1	10	0.20	0.79 (0.72, 0.85)	42 (38, 45)	−0.7 (−1.9, 0.4)
1	10	0.10	0.84 (0.77, 0.89)	170 (162, 177)	−1.3 (−1.9, −0.6)
1	33	0.05	0.88 (0.82, 0.92)	685 (673, 696)	−0.1 (−0.3, 0.1)
2	10	0.30	0.65 (0.57, 0.72)	13 (12, 14)	0.4 (−0.8, 1.8)
2	10	0.20	0.68 (0.60, 0.75)	17 (15, 19)	0.0 (−1.0, 1.1)
2	10	0.10	0.72 (0.64, 0.79)	44 (39, 48)	−0.5 (−1.2, 0.1)
2	10	0.05	0.80 (0.73, 0.86)	192 (179, 206)	−0.4 (−0.7, −0.1)
3	10	0.30	0.68 (0.60, 0.75)	38 (32, 44)	−2.5 (−4.6, −0.3)
3	10	0.20	0.65 (0.57, 0.72)	93 (79, 106)	−3.8 (−5.1, −2.5)
3	29	0.10	0.75 (0.67, 0.81)	534 (494, 575)	−1.3 (−2.0, −0.6)
3	119	0.05	0.92 (0.86, 0.95)	2469 (2396, 2543)	−0.1 (−0.4, 0.0)
4	10	0.30	0.65 (0.57, 0.72)	30 (26, 33)	3.0 (0.3, 5.8)
4	10	0.20	0.70 (0.62, 0.77)	45 (38, 52)	0.2 (−1.1, 1.5)
4	10	0.10	0.66 (0.58, 0.73)	113 (93, 133)	−2.1 (−2.7, −1.5)
4	36	0.05	0.81 (0.74, 0.87)	654 (594, 714)	−0.7 (−1.1, −0.4)
5	10	0.30	0.77 (0.69, 0.83)	30 (27, 32)	−0.9 (−2.6, 0.7)
5	10	0.20	0.84 (0.77, 0.89)	67 (62, 72)	−0.5 (−1.6, 0.5)
5	15	0.10	0.86 (0.79, 0.91)	307 (296, 318)	−0.8 (−1.3, −0.3)
5	61	0.05	0.87 (0.80, 0.92)	1250 (1232, 1268)	−0.2 (−0.5, 0.0)
6	10	0.30	0.72 (0.64, 0.79)	23 (21, 26)	0.4 (−1.1, 2.0)
6	10	0.20	0.67 (0.59, 0.74)	39 (34, 44)	−0.7 (−1.9, 0.4)
6	10	0.10	0.74 (0.66, 0.81)	131 (118, 143)	−0.3 (−1.1, 0.3)
6	32	0.05	0.87 (0.80, 0.92)	662 (640, 684)	−0.1 (−0.4, 0.0)
7	10	0.30	0.61 (0.53, 0.69)	72 (60, 84)	−7.1 (−9.0, −5.1)
7	15	0.20	0.67 (0.59, 0.74)	209 (183, 236)	−4.5 (−6.1, −2.8)
7	61	0.10	0.87 (0.80, 0.92)	1193 (1145, 1240)	−0.7 (−1.2, −0.2)
7	246	0.05	0.91 (0.85, 0.95)	5055 (4965, 5145)	0.0 (−0.2, 0.2)
8	10	0.30	0.62 (0.54, 0.70)	10 (10, 11)	−0.8 (−1.9, 0.3)
8	10	0.20	0.54 (0.46, 0.62)	90 (75, 106)	−3.4 (−4.6, −2.2)
8	32	0.10	0.73 (0.65, 0.80)	534 (472, 596)	−1.6 (−2.3, −0.9)
8	128	0.05	0.84 (0.77, 0.89)	2684 (2563, 2805)	−0.1 (−0.4, 0.2)
9	10	0.30	0.84 (0.77, 0.89)	27 (25, 29)	−0.8 (−2.2, 0.5)
9	10	0.20	0.83 (0.76, 0.88)	57 (54, 61)	−1.1 (−2.2, 0.0)
9	11	0.10	0.91 (0.85, 0.95)	242 (235, 248)	−0.7 (−1.2, −0.2)
9	47	0.05	0.90 (0.84, 0.94)	989 (976, 1002)	−0.2 (−0.5, 0.0)
10	10	0.30	0.77 (0.69, 0.83)	27 (25, 29)	−1.8 (−3.5, −0.1)
10	10	0.20	0.83 (0.76, 0.88)	62 (58, 66)	−1.5 (−2.5, −0.4)
10	12	0.10	0.86 (0.79, 0.91)	246 (238, 253)	−0.9 (−1.4, −0.4)
10	50	0.05	0.91 (0.85, 0.95)	1038 (1022, 1054)	0.0 (−0.2, 0.1)
11	10	0.30	0.82 (0.75, 0.87)	44 (40, 48)	0.2 (−1.4, 1.8)
11	10	0.20	0.84 (0.77, 0.89)	93 (86, 100)	−1.2 (−2.3, −0.1)
11	24	0.10	0.89 (0.83, 0.93)	425 (403, 447)	−0.8 (−1.3, −0.4)
11	98	0.05	0.93 (0.88, 0.96)	1944 (1890, 1998)	−0.2 (−0.5, 0.0)
12	10	0.30	0.81 (0.74, 0.87)	40 (35, 44)	−1.1 (−2.9, 0.6)
12	10	0.20	0.82 (0.75, 0.87)	89 (82, 96)	−1.6 (−2.7, −0.4)
12	18	0.10	0.94 (0.89, 0.97)	346 (329, 363)	−0.6 (−1.0, −0.1)
12	74	0.05	0.89 (0.83, 0.93)	1498 (1466, 1530)	−0.1 (−0.4, 0.0)

almost all experiments. In many experiments larger relative widths were adequate. For a fixed relative width, the expected number of tours varied widely from system to system. For $\gamma = 0.1$, the point estimates for the expected number of tours varied from 546 to 56435 for the M/G/1 queues. Thus, it would be extremely difficult to know ahead of time the fixed number of tours which should be simulated to achieve a desired relative width. However, for small sample sizes, fixed stopping rules yielded confidence intervals having more adequate coverage than those obtained using sequential stopping rules. Thus, if one needs only a rough but valid estimate of a response variable (i.e., if a valid confidence interval having an unspecified and possibly large relative width is adequate), then a fixed stopping rule may be appropriate. If, on the other hand, one would like a precise valid estimate of a response variable (i.e., if a valid confidence interval having a specified small relative width is desired), then a sequential stopping rule seems appropriate.

Since sequential stopping rules work poorly for large relative widths it is worthwhile to try to modify these rules to improve their small-sample properties (e.g.,

Table 5a Closed system with two types of customers: description of systems.

System	μ_{11}	μ_{12}	μ_{21}	μ_{22}	*State*	$p(1)$	$p(2)$	T
1	0.07	0.07	100.00	10.00	35 3 5 1	0.95	0.95	2.390
2	0.20	0.10	100.00	10.00	31 2 9 2	0.90	0.95	1.810

For all these cases $M = 6$, $\mu_3 = \mu_4 = \mu_5 = \mu_6 = 28.57$, $p_3 = p_4 = p_5 = p_6 = 0.25$, $N_1 = 40$, $N_2 = 4$.
For case 1 $\rho_2 = 0.82$, $\rho_3 = \rho_4 = \rho_5 = \rho_6 = 0.44$.
For case 2 $\rho_2 = 0.96$, $\rho_3 = \rho_4 = \rho_5 = \rho_6 = 0.57$.

Table 5b Closed system with two types of customers: empirical results.

System	K	γ	*Coverage*	*Number of Tours*	*Relative Bias (%)*
1	10	0.20	0.61 (0.53, 0.69)	50 (44, 56)	−0.2 (−2.1, 1.7)
1	10	0.10	0.70 (0.62, 0.77)	264 (244, 284)	−0.4 (−1.3, 0.5)
1	10	0.05	0.86 (0.76, 0.92)	1208 (1137, 1279)	−0.2 (−1.0, 0.5)
2	10	0.20	0.65 (0.57, 0.72)	15 (13, 17)	−0.1 (−1.2, 1.0)
2	10	0.10	0.75 (0.67, 0.81)	53 (49, 58)	−0.1 (−0.9, 0.5)
2	10	0.05	0.92 (0.83, 0.96)	253 (239, 267)	−0.2 (−0.6, 0.1)

[8]). In addition, it would be worthwhile to empirically compare the sequential stopping rules with a two-stage stopping rule in which the number of tours required to achieve a specified relative width is estimated during a short pilot run, and the run is then continued until the estimated number of tours have been simulated.

Appendix 1

Here we prove Lemmas 1–3.

Lemma 1

$\lim_{\gamma \to 0+} N(\gamma, \alpha) = \infty, \qquad P1.$

Proof It follows from (6) that for all $\omega \in \Omega$,

$$N(\gamma, \alpha, \omega) > [aS(N(\gamma, \alpha, \omega))/\gamma Y(N(\gamma, \alpha, \omega))]^2, \tag{11}$$

where $\alpha = 2\phi^{-1}[(1+\alpha)/2]$. Suppose

$$P(\lim_{\gamma \to 0+} N(\gamma, \alpha) = \infty) < 1.$$

Then, for some fixed finite integer N^*,

$$P(\lim_{\gamma \to 0+} N(\gamma, \alpha) = N^*) > 0.$$

Let $\Omega^* = \{\omega \in \Omega: \lim_{\gamma \to 0+} N(\gamma, \alpha, \omega) = N^*\}$.

Then for any $\omega \in \Omega^*$, there exists $\gamma_\omega^* > 0$, such that if $\gamma \leq \gamma_\omega^*$, then $N(\gamma, \alpha, \omega) = N^*$, $S(N(\gamma, \alpha, \omega)) = S(N^*) > 0$, and $Y(N(\gamma, \alpha, \omega)) = Y(N^*)$. Thus, from (11), if $\omega \in \Omega^*$, then

$$N^* > [aS(N^*)/Y(N^*)]^2/\gamma^2$$

for all $\gamma \leq \gamma_\omega^*$, where $aS(N^*)/Y(N^*)$ is a fixed positive number. This implies that $N^* = \infty$, a contradiction. Hence,

$$P(\lim_{\gamma \to 0+} N(\gamma, \alpha) = \infty) = 1. \qquad \square$$

Lemma 2 is a direct consequence of Lemma 1 and the strong law of large numbers; its proof is omitted.

Lemma 3

$\lim_{\gamma \to 0+} \gamma^2 N(\gamma, \alpha) = D, \qquad P1,$

where $D = \{2\phi^{-1}[(1+\alpha)/2]S/E[Y_1]\}^2$.

Proof It follows from (11) and Lemma 2 that

$$P(\liminf_{\gamma \to 0+} \gamma^2 N(\gamma, \alpha) \geq D) = 1.$$

It follows from (6) that for all $\omega \in \Omega$, if $S(N(\gamma, \alpha, \omega) - 1) > 0$, then

$$\gamma^2(N(\gamma, \alpha, \omega) - 1) \leq [aS(N(\gamma, \alpha, \omega) - 1)/Y(N(\gamma, \alpha, \omega) - 1)]^2.$$

Since

$$P(\lim_{\gamma \to 0+} S(N(\gamma, \alpha) - 1) = S) = 1,$$

where $S > 0$, it follows that for any $\epsilon > 0$ there exists $\gamma(\epsilon) > 0$ such that $P(S(N(\gamma, \alpha) - 1) > 0, \forall\, \gamma \leq \gamma(\epsilon)) \geq 1 - \epsilon$. Thus,

$$P(\gamma^2(N(\gamma, \alpha) - 1) \leq [aS(N(\gamma, \alpha) - 1)/Y(N(\gamma, \alpha) - 1)]^2, \quad \forall \gamma \leq \gamma(\epsilon)) \geq 1 - \epsilon.$$

Using Lemma 2, this implies that

$$P(\lim_{\gamma \to 0+} \sup \gamma^2 (N(\gamma, \alpha) - 1) \leq D) \geq 1 - \epsilon.$$

Since ϵ is arbitrary,

$$P(\lim_{\gamma \to 0+} \sup \gamma^2 N(\gamma, \alpha) \leq D) = 1,$$

which completes the proof. □

Appendix 2

We consider the cyclic queue described in the section on empirical studies. The ith cycle time T_i is the sum of two quantities, the time the customer spends in service center 1 during the cycle time, denoted $T_{i,1}$, and the time the customer spends in service center 2 during the cycle time, denoted $T_{i,2}$. The sequence $\{T_{i,1} : i = 1, 2, \cdots\}$ is a discrete parameter regenerative process with regeneration points $\{l_j : j = 1, 2, \cdots\}$, where l_j is the serial number of the jth arrival at service center 1 which finds the service center empty and leaves the process $\{\mathbf{V}(t) : t \geq 0\}$ in a fixed state. Since $\{\mathbf{V}(t) : t \geq 0\}$ is a finite-state irreducible Markov process, it follows that $\mathrm{E}[l_2 - l_1] < \infty$ and $\mathrm{E}[q_2 - q_1] < \infty$, where q_j is the time at which the l_jth arrival occurs. Furthermore, since there are N customers in the network,

$$\mathrm{E}\left[\sum_{i=l_1}^{l_2 - 1} T_{i,1}\right] \leq N \mathrm{E}[q_2 - q_1].$$

By using the above results it can be shown in a similar manner to that used in [13] that

$$T_1 = N_1 / \Lambda < \infty,$$

where Λ is given by (10), that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} T_{i,1} = T_1, \qquad P1,$$

and

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t n_1(u)\, du = N_1, \qquad P1.$$

The sequence $\{T_{i,2} : i = 1, 2, \cdots\}$ is a discrete parameter regenerative stochastic process with regeneration points $\{m_j : j = 1, 2, \cdots\}$, where m_j is the serial number of the jth arrival at service center 2 which finds the service center empty and leaves the process $\{\mathbf{V}(t) : t \geq 0\}$ in a fixed state. Thus, it can be shown that

$$T_2 = N_2 / \Lambda < \infty,$$

where

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} T_{i,2} = T_2, \qquad P1,$$

and

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t n_2(u)\, du = N_2, \qquad P1.$$

Table 6 M/G/1 queue: empirical results—fixed sampling.

System	*K*	*Coverage*	*Relative Bias (%)*
1	51	0.82 (0.75, 0.87)	−0.5 (−2.4, 1.1)
1	128	0.84 (0.77, 0.89)	−0.8 (−1.9, 0.3)
1	546	0.84 (0.77, 0.89)	−1.0 (−1.5, −0.5)
2	87	0.79 (0.72, 0.85)	−1.7 (−3.8, 0.3)
2	278	0.81 (0.74, 0.87)	−1.1 (−2.2, 0.0)
2	1498	0.89 (0.83, 0.93)	−0.7 (−1.2, −0.3)
3	305	0.81 (0.74, 0.87)	−2.7 (−4.6, −0.9)
3	1004	0.87 (0.80, 0.92)	−1.9 (−3.0, −0.8)
3	5325	0.88 (0.82, 0.92)	−0.9 (−1.3, −0.5)
4	207	0.86 (0.79, 0.91)	−1.2 (−2.9, 0.3)
4	520	0.84 (0.77, 0.89)	−1.6 (−2.7, −0.6)
4	2227	0.86 (0.79, 0.91)	−0.6 (−1.2, 0.0)
5	368	0.86 (0.79, 0.91)	−2.9 (−4.4, −1.4)
5	1083	0.94 (0.89, 0.97)	−0.7 (−1.7, 0.1)
5	4697	0.84 (0.77, 0.89)	−0.7 (−1.3, −0.2)
6	965	0.85 (0.78, 0.90)	−0.7 (−2.5, 0.9)
6	2742	0.88 (0.82, 0.92)	−1.3 (−2.3, −0.3)
6	12060	0.90 (0.84, 0.94)	−0.5 (−0.9, 0.0)
7	1799	0.89 (0.83, 0.93)	−0.8 (−2.3, 0.6)
7	4264	0.87 (0.80, 0.92)	−0.8 (−1.9, 0.1)
7	18167	0.87 (0.80, 0.92)	−0.3 (−0.8, 0.1)
8	3047	0.86 (0.79, 0.91)	−0.7 (−2.4, 0.7)
8	5583	0.91 (0.85, 0.95)	−1.3 (−2.3, −0.1)
8	31826	0.83 (0.76, 0.88)	−0.3 (−0.8, 0.2)
9	4967	0.86 (0.79, 0.91)	−2.0 (−3.5, −0.4)
9	12946	0.85 (0.78, 0.90)	−0.6 (−1.8, 0.4)
9	56435	0.89 (0.83, 0.93)	−0.5 (−1.0, 0.0)

Table 7 Closed system with two types of customers: empirical results—fixed sampling.

System	*K*	*Coverage*	*Relative Bias (%)*
2	15	0.76 (0.68, 0.82)	−0.3 (−1.4, 0.8)
2	53	0.85 (0.78, 0.90)	−0.1 (−0.8, 0.4)
2	253	0.82 (0.71, 0.89)	0.0 (−0.4, 0.3)

Thus, since $N_1 + N_2 = N$,

$$T = T_1 + T_2$$
$$= N / \Lambda.$$

Appendix 3

Here we give the method used to obtain confidence intervals for the coverage based on I independent replications. For the ith replication, let

$$\theta_i = \begin{cases} 1, & \text{if response variable contained in confidence interval,} \\ 0, & \text{else.} \end{cases}$$

Then $\theta_1, \cdots, \theta_I$ are iid random variables and $p = P(\theta_i = 1)$ is the true coverage. Let

$$\theta(I) = \frac{1}{I} \sum_{i=1}^{I} \theta_i.$$

Then,

$$\lim_{I\to\infty} P(I^{\frac{1}{2}}[\theta(I) - p]/[p(1-p)] \leq t) = \phi(t), \qquad (12)$$

and it can be shown that from (12) that for I sufficiently large, $(a(I, \beta) - b(I, \beta), a(I, \beta) + b(I, \beta))$ is approximately a $100 \times \beta$ percent confidence interval for p, where

$$a(I, \beta) = [\theta(I) + \psi^2(\beta)/2I]I/[I + \psi^2(\beta)],$$

$$b(I, \beta) = \{\theta(I)[1 - \theta(I)]/I + [\psi(\beta)/2I]^2\}^{\frac{1}{2}} \times I\psi(\beta)/[I + \psi^2(\beta)], \text{ and}$$

$$\psi(\beta) = \phi^{-1}[(1 + \beta)/2].$$

Appendix 4

For the simulation studies with the M/G/1 queues and central server models in the second set of experiments, we set K equal to $\max(10, D/20\gamma^2)$, where D, given in Lemma 3, is the value to which $\gamma^2 N(\gamma, \alpha)$ converges with probability one. We made this choice for the purpose of not having to compute the relative width of the estimated confidence interval more than about 20 times during a simulation run. We required that K be at least 10, since confidence intervals are estimated in APLOMB only if at least ten tours have been simulated. The value D was computed for the M/G/1 queues by using busy period analysis techniques (e.g., [18]); it was computed for the central server models using techniques for computing moments of first passage times in a semi-Markov process [19]. It is much more difficult to compute D for the cyclic queuing system and closed systems with two types of customers, and we did not attempt these computations. For these systems we set K equal to 10.

Acknowledgment

The authors thank B. Pittel for his contributions.

References

1. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems, I: General Multi-Server Queues," *J. ACM* **21,** 103 (1974).
2. M. A. Crane, and D. L. Iglehart, "Simulating Stable Stochastic Systems, II: Markov Chains," *J. ACM* **21,** 114 (1974).
3. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems, III: Regenerative Processes and Discrete-Event Simulations," *Oper. Res.* **23,** 33 (1975).
4. D. L. Iglehart, "Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators," *Naval Res. Logist. Quart.* **22,** 553 (1975).
5. D. L. Iglehart, "The Regenerative Method for Simulation Analysis," *Technical Report No. 86-20,* Control Analysis Corporation, Palo Alto, CA, 1975 (to appear in *Current Trends in Programming Methodology, Vol. 3, Software Modelling and Its Impact on Performance*, K. M. Chandy and R. T. Yeh, eds.
6. S. S. Lavenberg and D. R. Slutz, "Introduction to Regenerative Simulation," *IBM J. Res. Develop.* **19,** 458 (1975).
7. Y. S. Chow and H. Robbins, "On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean," *Ann. Math. Statist.* **36,** 457 (1965).
8. D. W. Robinson, "Determination of Run Length in Simulations of Stable Stochastic Systems," *Technical Report No. 86-21,* Control Analysis Corporation, Palo Alto, CA, 1976.
9. F. J. Anscombe, "Large-Sample Theory of Sequential Estimation," *Proc. Cambridge Phil. Soc.* **48,** 600 (1952).
10. C. H. Sauer, "Simulation Analysis of Generalized Queuing Networks," *Proceedings of the 1975 Summer Computer Simulation Conference*, San Francisco, 1975, p. 75.
11. K. L. Chung, *A Course in Probability Theory*, Harcourt, Brace and World, Inc., New York, 1968.
12. D. R. Cox and W. L. Smith, *Queues*, Methuen and Co., Ltd., London, 1961.
13. W. S. Jewell, "A Simple Proof of: $L = \lambda W$," *Oper. Res.* **15,** 1109 (1967).
14. J. P. Buzen, "Analysis of System Bottlenecks Using a Queueing Network Model," *ACM-SIGOPS Workshop on System Performance Evaluation*, ACM, New York, 1971, p. 82.
15. L. Kleinrock, *Queueing Systems Volume 2: Computer Applications*, John Wiley & Sons, Inc., New York, 1976.
16. M. Reiser and H. Kobayashi, "Queuing Networks with Multiple Closed Chains: Theory and Computational Algorithms," *IBM J. Res. Develop.* **19,** 283 (1975).
17. P. A. W. Lewis, A. S. Goodman, and J. M. Miller, "Pseudo-Random Number Generator for the System/360," *IBM Syst. J.* **8,** 136 (1969).
18. A. M. Law "Efficient Estimators for Simulated Queuing Systems," *Technical Report ORC 74-7*, Operations Research Center, University of California, Berkeley, 1974.

Received February 28, 1977

S. S. Lavenberg is located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598; C. H. Sauer is located at the Department of Computer Sciences, University of Texas, Austin, Texas 78712.