Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

MATHEMATICS FOR SIMULATION

Shane G. Henderson

Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109-2117, U.S.A.

ABSTRACT

I survey several mathematical techniques and results that are useful in the context of stochastic simulation. The concepts are introduced through the study of a simple model of ambulance operation to ensure clarity, concreteness and cohesion.

1 INTRODUCTION

The incredibly rich set of mathematical tools and techniques that underlie stochastic simulation is the subject of this paper. Of course, the field is far too large to be covered in a single paper, and so I choose to focus the discussion somewhat. For example, there is no discussion in this paper on the vast array of techniques that may be used for input analysis, uniform and nonuniform random variate generation, sensitivity analysis, and so forth. For excellent overviews of these and other topics, see Bratley, Fox and Schrage (1987), Law and Kelton (2000), and the tutorials and advanced tutorials in recent proceedings of the Winter Simulation Conference.

Instead, what I attempt to do is to describe a set of mathematical tools and techniques that can be used to rigorously define performance measures in both the terminating and steady-state context. I have also attempted to describe methods that can be used to shed light on the properties of simulation-based estimators of these measures.

The emphasis is always on the mathematical results and techniques that can be used to derive the results. It would be very easy to provide a smorgasbord of such results, but such a paper would read like an encyclopedia. Therefore, I introduce a simple model of ambulance operation that serves to unify the discussion, and define several performance measures related to this model.

All of the performance measures described in this paper take the form of an expectation of a random variable, or a differentiable function of a finite number of expectations. Such performance measures are useful when the goal is to compare many different stochastic systems, as they provide a concrete basis for the comparison. However, if the goal is to enhance one's *understanding* of a single stochastic system, then it is often more useful to analyze the *distribution* of certain random variables, perhaps through density estimation techniques. Unfortunately, there is not space in this article to delve into specific techniques in the theory of density estimation. This is a shame, since density estimation is bound to become a more important area of research as users of simulation become more sophisticated, and the theory matures. Nevertheless, many of the techniques presented here can be applied in such contexts; see Henderson and Glynn (1999), for example.

In Section 2 I introduce the essential elements of the ambulance model that serves as the underlying thread of the paper. This then sets the stage for the next two sections of the paper, which specialize the model to first the terminating simulation context, and second, the steady-state context.

In Section 3 we find ourselves in the terminating simulation context, in which there is a finite time interval over which the simulation will be run. One might then be interested in performance measures like the expected utilization of the ambulance, or the expected response time to calls. The primary tool in defining measures such as these is the strong law of large numbers, which also motivates several estimators of the performance measures.

Of course, in any numerical analysis method, and simulation is certainly one such method, it is important to provide error bounds. Such error bounds can be derived through the central limit theorem. To apply the central limit theorem to yield confidence intervals, several constants must be replaced with sample estimates, and one should question whether the resulting confidence intervals are then valid. One approach to establishing this validity is via the continuous mapping theorem.

The last key idea in Section 3 is one that is often useful in simulation analysis. Some performance measures cannot be written as the expectation of a random variable, but may be written as a differentiable function of certain expectations. In this setting, Taylor's theorem is a very useful tool that can be used to derive central limit theorems that then form the basis for confidence interval construction.

One can also examine the bias properties of estimators in this context using arguments based on Taylor's theorem.

By imposing different assumptions on the model, one obtains a steady-state simulation, where the performance measures are all long-run averages. To rigorously define these performance measures, it is necessary to define an appropriate stochastic process with which to work. A great deal is known about the class of Markov processes evolving on general (not necessarily countable) state spaces. In Section 4 a general state space Markov chain is defined. To ensure that long-run averages exist, it is necessary to show that this chain is, in a certain sense, positive recurrent.

A very practical approach to establishing that a Markov chain is positive recurrent is to use Lyapunov functions, and this approach is the central mathematical tool illustrated in Section 4. We use Lyapunov theory to show that certain Markov chains are positive recurrent, that our performance measures are well-defined, that certain estimators are consistent and satisfy central limit theorems, and that confidence intervals obtained through the method of batch means are asymptotically valid. An important consideration in the steady-state context is that of initialization bias. We also use Lyapunov theory to characterize the magnitude of such bias.

The underlying theme of Section 4 is then that Lyapunov functions provide an enormously powerful, and easily applied (at least relative to many other methods!) approach to establishing results that underlie steady-state simulation methodology.

Throughout this paper, results are rigorously quoted, and references given for the proofs. To simplify the exposition, it is often the case that results are quoted using stronger hypotheses than are strictly necessary, but tighter hypotheses can be found in the references provided.

2 A SIMPLE MODEL

To begin, we describe a very simple model that will serve as a vehicle for the concepts to follow. The purpose of the example is therefore simplicity, and certainly not realism, although with a few straightforward extensions, the model could be considered to be quite practical.

Suppose that a single ambulance serves calls in a square region. By translating and rescaling units, we may assume that the square is centred at the origin, with lower left-hand corner at $(-1/2, -1/2)$ and upper right-hand corner at $(1/2, 1/2)$. For simplicity, we assume that the ambulance travels at unit speed within the square. The combined hospital/ambulance base is located at the origin.

Calls arrive (in time) according to a homogeneous Poisson process with rate λ. The location of the call is independent of the arrival process, and uniformly distributed over the square. To serve a call, the ambulance travels in a Manhattan fashion (i.e., at any given time, movement is restricted to lie only in the x direction or the y direction) from its present location to the location of the call. A random amount of time is then spent at the scene treating the patient, independent of all else. After this scene time is complete, with probability p (independent of all else), the ambulance is required to transport and admit the patient to the hospital, with hospital admission occurring instantaneously once the ambulance reaches the hospital, and with probability $1 - p$ the ambulance is freed for other work.

3 TERMINATING SIMULATION

In this section, we assume that the ambulance only receives calls from (say) 7am until 11pm each day. At 11pm, the ambulance completes the call that it is currently serving (if any) and returns to base. We will further assume that if the ambulance is engaged with a call when another call is received, then some outside agency, such as another emergency service, handles the other call. Finally, we assume that the random variables associated with each day are independent of those for all other days.

We will be interested in several performance measures relating to ambulance operation as follows.

- α_1 The long-run utilization of the ambulance, i.e., the percentage of time that the ambulance is occupied with a call.
- α_2 The long-run fraction of calls attended by the ambulance.
- α_3 The long-run fraction of calls with response time (time from when the call arrives to when the ambulance arrives at the scene) being at most t^* time units, where the fraction does not take into account those calls that are handled by the outside agency.
- α_4 The long-run average response time for those calls handled by the ambulance, and not by the outside agency.

In order to more carefully define these performance measures, we proceed as follows. Let T_i denote the total number of hours that the ambulance is busy on day i, with $T_i \leq 16$, since we will not count any residual time after 11pm needed to complete any call in progress. After n days then, the average utilization of the ambulance is

$$\alpha_1(n) = \frac{\sum_{i=1}^{n} T_i}{16n},$$

and the long-run utilization α_1 is the limiting value of $\alpha_1(n)$. To ensure that α_1 is properly defined, we need to ensure that this limit exists, and is the same, regardless of the particular realization $T_1, T_2, \ldots$ involved. Observe that $(T_n : n \geq 1)$ is an i.i.d. sequence of bounded random variables, and so the strong law of large numbers (SLLN) is applicable.

Theorem 1 (SLLN) *If $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with $E|X_1| < \infty$, then*

$$\frac{\sum_{i=1}^n X_i}{n} \to EX_1 \text{ a.s.}$$

as $n \to \infty$.

For a proof, see p. 290 of Billingsley (1986).

The SLLN implies that as $n \to \infty$, the limit of $\alpha_1(n)$ exists and uniquely defines α_1.

In view of the SLLN, a natural estimator of α_1 is $\alpha_1(n)$, and the SLLN ensures that this estimator converges to α_1 as $n \to \infty$ almost surely, i.e., that $\alpha_1(n)$ is consistent.

Now let us consider α_2, the long-run fraction of calls attended by the ambulance. Let N_i denote the total number of calls received on day i, and let A_i denote the number of those calls attended by the ambulance. After n days, the fraction of calls attended by the ambulance is given by

$$\frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n N_i}. \tag{1}$$

Dividing both the numerator and denominator of (1) by n, and applying the SLLN separately to both the numerator and denominator, we see that

$$\frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n N_i} \to \alpha_2 = \frac{EA_1}{EN_1} \text{ a.s.}$$

as $n \to \infty$. Note that EN_1, the expected number of calls received in one day is known, and equal to 16λ. Thus, one can estimate α_2 by

$$\alpha_2(n) = \frac{\sum_{i=1}^n A_i}{16\lambda n}.$$

But how can one assess the accuracy of the estimators $\alpha_1(n)$ and $\alpha_2(n)$? One answer is via the central limit theorem (CLT).

Theorem 2 (CLT) *If $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with $EX_1^2 < \infty$, then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - EX_1\right) \Rightarrow \sigma N(0,1)$$

as $n \to \infty$, where $\sigma^2 = \text{var}\, X_1$, $\Rightarrow$ denotes weak convergence, and $N(0,1)$ denotes a standard normal random variable.

For a proof, see p. 367 of Billingsley (1986).

The CLT basically establishes that the error in the estimator $\alpha_1(n)$ is approximately normally distributed with mean 0 and variance $s^2/256n$ where $s^2 = \text{var}\, T_1$, and this is the basis for obtaining confidence intervals for α_1. In particular, an approximate 95% confidence interval for α_1 is given by

$$\alpha_1(n) \pm 1.96\sqrt{\frac{s^2}{256n}}. \tag{2}$$

However, s^2 must invariably be estimated. The usual estimator is the sample variance

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\sum_{i=1}^n (T_i - \bar{T}_n)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n T_i^2 - \frac{n}{n-1}\bar{T}_n^2, \end{aligned}$$

where $\bar{T}_n = n^{-1}\sum_{i=1}^n T_i$ is the usual sample mean. The confidence interval that is reported is the same as (2) with s^2 replaced with its sample counterpart s_n^2. But is the modified confidence interval then valid?

The SLLN implies that $s_n^2 \to s^2$ a.s. as $n \to \infty$. Hence, by Exercise 29.4 of Billingsley (1986), we have that

$$\begin{pmatrix} n^{1/2}(\bar{T}_n - ET_1) \\ s_n^2 \end{pmatrix} \Rightarrow \begin{pmatrix} sN(0,1) \\ s^2 \end{pmatrix}. \tag{3}$$

The natural tool to apply at this point is the continuous mapping theorem. For a real-valued function h in $\mathbf{R}^d$, let D_h denote its set of discontinuities (in $\mathbf{R}^d$).

Theorem 3 (Continuous Mapping Theorem) *Let $(X_n : n \geq 1)$ be a sequence of $\mathbf{R}^d$ valued random variables with $X_n \Rightarrow X$ as $n \to \infty$ and let $h : \mathbf{R}^d \to \mathbf{R}$ be measurable. If $P(X \in D_h) = 0$, then $h(X_n) \Rightarrow h(X)$ as $n \to \infty$.*

For a proof, see p. 391 of Billingsley (1986).

Define $h(x, y) = x/y^{1/2}$, and then apply the continuous mapping theorem to (3), to obtain that

$$\frac{n^{1/2}(\bar{T}_n - ET_1)}{s_n} \Rightarrow N(0,1)$$

as $n \to \infty$, and so the confidence interval procedure outlined above is indeed valid.

The analysis for the estimator $\alpha_2(n)$ of α_2 is similar.

Let us now consider the performance measures α_3 and α_4. First consider α_3.

Recall that on day i, the ambulance responds to A_i calls out of a possible N_i, with the remainder being served by some outside agency. For $1 \leq j \leq A_i$, let R_{ij} denote the response time for the jth call handled by the ambulance

on day i. Let $I(R_{ij} \leq t^*)$ denote the indicator random variable that is 1 if $R_{ij} \leq t^*$, and 0 otherwise. Then

$$Y_i = \sum_{j=1}^{A_i} I(R_{ij} \leq t^*)$$

denotes the number of calls that the ambulance reached in at most t^* time units on day i. Over the first n days, the fraction of calls with response time at most t^* is then

$$\alpha_3(n) = \frac{n^{-1}\sum_{i=1}^n Y_i}{n^{-1}\sum_{i=1}^n A_i}. \tag{4}$$

and the strong law of large numbers implies that $\alpha_3(n)$ converges almost surely to $\alpha_3 = EY_1/EA_1$.

So how can one assess the accuracy of the estimator $\alpha_3(n)$? Certainly, the standard central limit theorem cannot be applied, because $\alpha_3(n)$ is a *ratio* of sample means of i.i.d. observations. We first consider a strongly related question, and then return to the problem at hand.

Suppose that $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with finite mean $\mu = EX_1$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ denote the sample mean. If h is continuous at μ, it follows that $h(\bar{X}_n) \to h(\mu)$ a.s. as $n \to \infty$. So how does the error $h(\bar{X}_n) - h(\mu)$ behave, for large n? Note that for large n, $\bar{X}_n$ will be very close to μ, and so the asymptotic behaviour of the error should depend only on the local behaviour of h. Indeed, if h is appropriately differentiable, then Taylor's theorem implies that

$$h(\bar{X}_n) - h(\mu) \approx h'(\mu)(\bar{X}_n - \mu),$$

and so if the X_i's have finite variance, then

$$\begin{aligned} n^{1/2}(h(\bar{X}_n) - h(\mu)) &\approx h'(\mu)n^{1/2}(\bar{X}_n - \mu) \\ &\Rightarrow \eta N(0,1) \end{aligned}$$

as $n \to \infty$, where $\eta^2 = h'(\mu)^2 \operatorname{var} X_1$.

This intuitive argument can be formalized, and also generalized to higher dimensions to obtain the following result, sometimes referred to as the delta method.

Theorem 4 *Suppose that $(X_n : n \geq 1)$ is an i.i.d. sequence of $\mathbb{R}^d$ valued random variables with $E\|X_1\|_2^2 < \infty$. Let $\mu = EX_1$ denote their common mean, and let Λ denote their common covariance matrix. Let $\bar{X}_n$ denote the sample mean of $X_1, \ldots, X_n$. If $h : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable in a neighbourhood of μ with non zero gradient at μ, then*

$$n^{1/2}(h(\bar{X}_n) - h(\mu)) \Rightarrow \sigma N(0,1)$$

as $n \to \infty$, where $\sigma^2 = g'\Lambda g$, and $g = \nabla h(\mu)$.

For a proof, see p. 122 of Serfling (1980).

To apply this result in our context, let $X_i = (Y_i, A_i)$, and define $g(y,a) = y/a$. We then find that

$$n^{1/2}(\alpha_3(n) - \alpha_3) \Rightarrow \sigma N(0,1),$$

where

$$\sigma^2 = \frac{E(Y_1 - \alpha_3 A_1)^2}{(EA_1)^2}.$$

Using the SLLN, one can easily show that σ^2 can be consistently estimated by

$$s_n^2 = \frac{n^{-1}\sum_{i=1}^n (Y_i - \alpha_3(n)A_i)^2}{(n^{-1}\sum_{i=1}^n A_i)^2},$$

and the same continuous mapping argument used for the estimators $\alpha_1(n)$ and $\alpha_2(n)$ establishes that

$$\alpha_3(n) \pm 1.96 s_n/\sqrt{n}$$

is an approximate 95% confidence interval for α_3.

The estimator

$$\alpha_4(n) = \frac{\sum_{i=1}^n \sum_{j=1}^{A_i} R_{ij}}{\sum_{i=1}^n A_i}$$

can be handled in exactly the same fashion.

Taylor's theorem can also be used to examine the bias properties of the estimators $\alpha_3(n)$ and $\alpha_4(n)$. In particular, using our previous notation, Taylor's theorem implies that

$$h(\bar{X}_n) - h(\mu) \approx h'(\mu)(\bar{X}_n - \mu) + \frac{1}{2}h''(\mu)(\bar{X}_n - \mu)^2.$$

Taking expectations, we find that

$$Eh(\bar{X}_n) - h(\mu) \approx \frac{1}{2}h''(\mu) \operatorname{var} X_1/n,$$

i.e., we have an explicit expression for the asymptotic bias. As before, this argument can be formalized, and generalized to higher dimensions.

Theorem 5 *Suppose that $(X_n : n \geq 1)$ is an i.i.d. sequence of $\mathbb{R}^d$ valued random variables with $E\|X_1\|_2^4 < \infty$. Let $\mu = EX_1$ denote their common mean, and let Λ denote their common covariance matrix. Let $\bar{X}_n$ denote the sample mean of $X_1, \ldots, X_n$. If $h : \mathbb{R}^d \to \mathbb{R}$ is such that $h(\bar{X}_n)$ is bounded for all n with probability 1, and twice continuously differentiable in a neighbourhood of μ, then*

$$n(Eh(\bar{X}_n) - h(\mu)) \to \frac{1}{2}\sum_{i,j=1}^d \nabla^2 h(\mu)_{ij}\Lambda_{ij}$$

as $n \to \infty$.

The proof is a slight modification of Theorem 7 in Glynn and Heidelberger (1990).

We would like to apply this result to both of the estimators $\alpha_3(n)$ and $\alpha_4(n)$. The only condition that is not obviously satisfied is that $h(\bar{X}_n)$ is bounded for all n with probability 1. In both cases, we take $h(x, y) = x/y$. Note that $\alpha_3(n) = h(\bar{Y}_n, \bar{A}_n) \leq 1$. For $\alpha_4(n)$, observe that the sum of the response times on any day is bounded by 16 hours, plus any response time that carries over the end of the 16 hour day. Since the ambulance takes at most 2 hours to travel from anywhere in the square to anywhere else, the sum of the response times on any day is bounded by 18. This then allows us to conclude that $\alpha_4(n)$ is also bounded with probability 1.

We have therefore established that the bias in the estimators $\alpha_3(n)$ and $\alpha_4(n)$ is of the order n^{-1}.

It is reasonable to ask whether this bias is sufficient to noticeably affect the performance of the confidence intervals produced earlier for a given runlength n. Recall that the widths of the confidence intervals are of the order $n^{-1/2}$. Thus, the bias decreases at a (much) faster asymptotic rate than the width of the confidence intervals, and so unless runlengths are quite small, it is reasonable to neglect bias.

4 STEADY-STATE SIMULATION

We now turn to useful mathematical techniques and results for steady-state simulation analysis. For this purpose, we will modify the assumptions of the previous section on the dynamics of the ambulance model. In particular, in addition to the assumptions given in Section 2, we assume that the ambulance operates 24 hours a day, 7 days a week. Furthermore, calls that arrive while the ambulance is busy are queued, and answered in first-in first-out order. Once the current call is complete, the ambulance then attends to the next call. Recall that a call is completed either at the scene (with probability $1 - p$), or when the ambulance drops the patient off at the hospital (with probability p).

For this model, 3 of the previous 4 performance measures are still relevant, but because the ambulance is now handling *all* calls, the fraction of calls answered by the ambulance (α_2) is no longer of interest. For convenience, and also to refine the statement of the performance measures to our new setting, we restate the performance measures.

- β_1 The long-run utilization of the ambulance, i.e., the percentage of time that the ambulance is occupied with a call.
- β_2 The long-run fraction of calls with response time being at most t^* time units.
- β_3 The long-run average response time.

In the previous section we attempted to rigorously define the suggested performance measures, and also to derive asymptotic results that lay at the heart of confidence interval methodology for estimating them. We will proceed in a similar fashion in this section. The 3 performance measures given above all involve the term "long-run". In order that such long-run measures exist, it is first necessary that the ambulance model be stable. In order to be able to make statements about the stability, or lack thereof, of the model, it is first necessary to define an appropriate stochastic process from which our performance measures can be derived. Statements about the stability of the model really relate to the stability of the stochastic process.

There are typically a host of stochastic processes that may be defined from the elements of a simulation. The choice of stochastic process depends partly on the performance measures in question. Given that two of our measures are related to response time, it is natural to consider a stochastic process that yields information on response times. Furthermore, for mathematical convenience, it is often helpful to ensure that one's stochastic process is Markov.

For $n \geq 1$, let T_n denote the time at which the nth call is received, with $T_0 = 0$. For $n \geq 1$, let W_n be the *residual workload* of the ambulance at time T_n+, i.e., immediately after the nth call is received. By residual workload at some time t, we mean the amount of time required for the ambulance to complete any current call, along with calls that might also be queued at time t. We *define* $W_0 = 0$.

Unfortunately, $(W_n : n \geq 0)$ is not a Markov process, because the response time for a future call, and hence the workload, depends on the location of the ambulance when the ambulance clears the previous workload. So if we also keep track of the location coordinates of the ambulance (X_n, Y_n) at the instant at which the workload W_n is first cleared, then the resulting process $Z = (Z_n : n \geq 0)$ is Markov, where $Z_n = (W_n, X_n, Y_n)$. We define $X_0 = Y_0 = 0$, i.e., the ambulance begins at the hospital.

The process Z is a general state space Markov chain, and evolves on the state space

$$S = [0, \infty) \times [-1, 1]^2.$$

The first step in ensuring that our "long-run" performance measures are defined is to establish that Z exhibits some form of positive recurrence. One way to achieve this is to verify that the chain Z satisfies the following condition, which certainly deserves some explanation!

To avoid a potential confusion between general results and those for our particular model, we will state general results in terms of a Markov chain $\Phi = (\Phi_n : n \geq 0)$ evolving on a state space $\mathcal{S}$.

The Lyapunov Condition There exists a $B \subseteq \mathcal{S}$, positive scalars $a < 1, b$, and δ, an integer $m \geq 1$, a probability distribution φ on $\mathcal{S}$, and a function $V : \mathcal{S} \to [1, \infty)$ such that

1. $P(\Phi_m \in \cdot | \Phi_0 = z) \geq \delta\varphi(\cdot)$ for all $z \in B$, and
2. $E(V(\Phi_1)|\Phi_0 = z) \leq aV(z) + bI(z \in B)$ for all $z \in \mathcal{S}$.

The Lyapunov condition (sometimes called a Foster-Lyapunov condition) is a stronger condition than we really require, but it simplifies the presentation considerably. The function V is called a Lyapunov (think of energy) function. The second requirement basically states that when the chain Φ lies outside of the set B, the energy in the system tends to decrease, and when the chain lies inside B, the energy in the system cannot become too big on the next step. This condition implies that the set B gets hit infinitely often. Of course, if one takes $B = \mathcal{S}$, the entire state space, then this requirement is trivially satisfied. The first condition is needed to ensure that the set B is not too "big".

In any case, the point is that if a chain Φ satisfies the Lyapunov condition, then Φ is appropriately positive recurrent. The precise statement is as follows.

Theorem 6 *If a discrete time Markov chain Φ is aperiodic and satisfies the Lyapunov condition, then it is V-uniformly ergodic. In particular, Φ has a unique stationary probability distribution.*

For a proof, see Theorem 16.0.1 of Meyn and Tweedie (1993).

So the question then is, does our chain Z satisfy the Lyapunov condition? The answer is yes, and it is instructive to go through a proof. However, on a first reading one may skip the following development up to the statement of Proposition 7 without loss of continuity.

For many systems, the function V may be taken to be $e^{\gamma v}$, where v is some measure of the work in the system. In fact, as we now show, one may take $V(w, x, y) = e^{\gamma w}$ for some yet to be determined constant $\gamma > 0$.

Consider what happens on a single transition of the chain Z, starting from the point (w, x, y). There will be some delay, τ say, until the next call is received, and during this time the workload decreases at unit rate, at least until it hits zero. At the instant that the new call arrives, we add the time η_1 required for the ambulance to travel to the new call. We also add the time required to treat the patient at the scene, U say. A Bernoulli random variable ξ with $P(\xi = 1) = p$, indicates whether the patient needs transport to the hospital ($\xi = 1$), or not ($\xi = 0$). If $\xi = 1$, then the workload also includes the travel time η_2 to the hospital.

In summary then, the new workload W_1 is given by

$$W_1 = [w - \tau]^+ + \eta_1 + U + \xi\eta_2,$$

where $[x]^+ = \max\{x, 0\}$, and τ, U and ξ are independent of each other and of (η_1, η_2).

So if $z = (w, x, y)$, then $E[V(Z_1)|Z_0 = z]$ is given by

$$\begin{aligned}
& E(e^{\gamma W_1}|Z_0 = (w, x, y)) \\
= \;& Ee^{\gamma(\eta_1+U+\xi\eta_2)}Ee^{\gamma[w-\tau]^+} \\
\leq \;& Ee^{\gamma(\eta_1+U+\xi\eta_2)}\,(Ee^{\gamma(w-\tau)} + P(w - \tau \leq 0)) \\
\leq \;& e^{\gamma w}Ee^{\gamma(\eta_1+U+\xi\eta_2-\tau)}\,(1 + Ee^{\gamma(\tau-w)}) \\
\leq \;& e^{\gamma w}Ee^{\gamma(2+U+1-\tau)}\,(1 + e^{-\gamma w}Ee^{\gamma\tau}) \qquad (5) \\
= \;& V(w)\phi(\gamma)(1 + e^{-\gamma w}Ee^{\gamma\tau}),
\end{aligned}$$

where the function ϕ is defined appropriately. Equation (5) follows since the ambulance travels at unit rate, and the distances it can travel are such that $\eta_1 \leq 2$, and $\eta_2 \leq 1$. (Recall that the ambulance travels distances as measured by the Manhattan metric.)

Assuming that $e^{\gamma U}$ is finite in a neighbourhood of 0, i.e., U has a moment generating function defined near 0, then we have that $\phi(0) = 1$, and

$$\phi'(0) = E(U + 3 - \tau).$$

So if $EU + 3 < 1/\lambda$, then $\phi'(0) < 0$, and so $\phi(\gamma) < 1$ for γ in some neighbourhood of 0. Now, we also require that $Ee^{\gamma\tau} < \infty$, which is true for $\gamma < \lambda$, since τ has an exponential distribution with rate λ. So choose $\gamma \in (0, \lambda)$ so that $\phi(\gamma) < 1$. We then have that

$$E[V(Z_1)|Z_0 = z] \leq V(w)\phi(\gamma)(1 + e^{-\gamma w}Ee^{\gamma\tau}).$$

Now, there is some $K > 0$ such that if $w > K$, then

$$\phi(\gamma)(1 + e^{-\gamma w}Ee^{\gamma\tau}) < 1,$$

since $Ee^{\gamma\tau} < \infty$ and $\phi(\gamma) < 1$. Furthermore, for $w \leq K$ we have that

$$E[V(Z_1)|Z_0 = z] \leq e^{\gamma(K+3+U)} < \infty.$$

Thus, if we take $B = [0, K] \times [-1, 1]^2$, then the second requirement in the Lyapunov condition is met.

It remains to check the first requirement. Observe that if the time till the next call is large enough, then the ambulance will have reached its base after serving all calls. In particular, if $\tau > K + 1$, then independent of $z \in B$, the next call will be served immediately by the ambulance from the base. If we let φ denote the distribution of Z_1 under this scenario, then we immediately have that for all $z \in B$,

$$P(z, \cdot) \geq e^{-\lambda(K+1)}\varphi(\cdot),$$

and the first requirement in the Lyapunov condition is satisfied.

In summary then, we have established that Z satisfies the Lyapunov condition. It is straight-forward to show that

Z is aperiodic, and so we arrive at the following result. Recall that U is a generic service time (time spent at the scene), and τ represents a generic interarrival time.

Proposition 7 *If U possesses a moment generating function in a neighbourhood of 0, and $EU+3 < E\tau$, then the chain Z is V-uniformly ergodic, where $V(w,x,y) = e^{\gamma w}$, for some $\gamma > 0$.*

The stability condition

$$EU + 3 < E\tau$$

has a very nice interpretation in terms of the model. The left-hand side of the inequality gives an upper bound on the expected amount of work (time at the scene + travel time to the scene + travel time from the scene to the hospital) brought in by an arriving call, whereas the right-hand side gives the expected amount of time that the ambulance has between calls to deal with this work. This condition can certainly be weakened by being more careful about defining how much work each call brings to the system, but this is not something that we will pursue further.

The main point is that Proposition 7 gives *easily verifiable* conditions under which the system is stable. While it may have appeared somewhat difficult to verify the Lyapunov condition, the argument used is actually quite straightforward, and we will see that the payoff is easily worth the effort. Based on this result, we can now define our performance measures rigorously, and also construct estimators that we can prove are consistent and satisfy central limit theorems.

As in Section 3, the rigorous definition of all of our performance measures is based on the strong law of large numbers. For simplicity, we state this theorem under stronger hypotheses than are really necessary.

Theorem 8 (MCSLLN) *Suppose that Φ is a V-uniformly ergodic Markov chain on state space $\mathcal{S}$ with stationary probability distribution π. Let $f : \mathcal{S} \to \mathbb{R}$ be a real-valued function on $\mathcal{S}$. If $\pi|f| = \int_{\mathcal{S}} |f(x)|\pi(dx) < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n-1} f(\Phi_i) \to \pi f \text{ a.s.}$$

as $n \to \infty$.

For a proof, see Theorem 17.0.1 of Meyn and Tweedie (1993).

Let us return now to the performance measures we outlined earlier. First, consider β_1, the utilisation of the ambulance. The actual utilisation of the ambulance over the time interval $[0, T_n)$, i.e., up until the time of the nth arrival is

$$\frac{n^{-1}\sum_{i=0}^{n-1} \min\{W_i, \tau_{i+1}\}}{n^{-1}\sum_{i=0}^{n-1} \tau_{i+1}}, \tag{6}$$

where, for $i \geq 0$, τ_{i+1} denotes the time between the ith and $(i+1)$th arrival. Now, the SLLN for i.i.d. random variables implies that the denominator converges to λ^{-1}. We would like to apply the MCSLLN to the numerator, but it is not yet in an appropriate form. However, using a simple device we can fix this difficulty. In essence, we are going to apply filtering; see Glasserman (1993). We have that

$$\begin{aligned} E\min\{w, \tau_1\} &= wP(\tau_1 > w) + E\tau_1 I(\tau_1 \leq w) \\ &= \lambda^{-1}(1 - e^{-\lambda w}), \end{aligned}$$

and so we replace (6) by

$$\beta_1(n) = \frac{1}{n}\sum_{i=0}^{n-1}(1 - e^{-\lambda W_i}). \tag{7}$$

Notice that $\beta_1(n)$ is in exactly the form that we need to apply the MCSLLN, with $f(w,x,y) = 1 - e^{-\lambda w}$, which is bounded, and so we find that

$$\beta_1(n) \to \beta_1 \text{ a.s.}$$

as $n \to \infty$. This then is a rigorous definition of β_1, and also a proof that the estimator $\beta_1(n)$ is (strongly) consistent.

Turning now to the performance measures β_2 and β_3, first note that both measures are related to the response times of the ambulance to calls. The response time R_n of the ambulance to the nth call is given, for $n \geq 1$, by

$$\begin{aligned} R_n &= [W_{n-1} - \tau_n]^+ + \eta_1(n) \\ &= W_n - U_n - \xi_n\eta_2(n), \end{aligned}$$

where W_n is the workload just after the nth call arrives, $\eta_1(n)$ is the time required for the ambulance to travel to the location of the nth call, U_n is the service time at the nth call, ξ_n is the indicator variable that is 1 if the patient needs to be transported to hospital, and $\eta_2(n)$ is the time required for the ambulance to travel to the hospital from the location of the nth call.

Observe that we cannot write R_n as a (deterministic) function of Z_n. We could apply a filtering method as above, but it is instructive to adopt a different approach. Suppose that we append additional information to the process Z, creating a new Markov chain $\tilde{Z} = (\tilde{Z}(n) : n \geq 0)$. In particular, let

$$\tilde{Z}(n) = (W_n, X_n, Y_n, U_n, \xi_n).$$

Using the same methods as before, we can show that $\tilde{Z}$ is $\tilde{V}$-uniformly ergodic and aperiodic, where $\tilde{V}$ is the function

$$\tilde{V}(w,x,y,u,\xi) = e^{\gamma w}.$$

We now have that $R_n = r(\tilde{Z}_n)$ say, where the function $r(\cdot)$ is defined by

$$r(w, x, y, u, \xi) = w - u - \xi d((x, y), (0, 0))$$

and d is the function returning the (Manhattan) distance between its two arguments.

The fraction of the first n calls for which the response time is less than t^* is

$$\beta_2(n) = \frac{1}{n}\sum_{i=1}^{n} I(R_i \le t^*), \tag{8}$$

and the MCSLLN immediately implies that $\beta_2(n) \to \beta_2$ a.s. as $n \to \infty$, thus both defining β_2 and proving that the estimator $\beta_2(n)$ is consistent.

The mean response time over the first n calls is

$$\beta_3(n) = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} r(\tilde{Z}_i). \tag{9}$$

To apply the MCSLLN, we need to show that if $\tilde{\pi}$ is the stationary distribution of $\tilde{Z}$, then $\tilde{\pi} r < \infty$. The following result is extremely useful in this regard.

Proposition 9 *Suppose that the Lyapunov condition holds for a Markov chain Φ on state space $\mathcal{S}$ with stationary probability distribution π. Then for any function $f : \mathcal{S} \to \mathbb{R}$ with $|f(z)| \le V(z)^{1/2}$ for all $z \in \mathcal{S}$, $\pi f < \infty$.*

For a proof, see Lemma 17.5.1 of Meyn and Tweedie (1993).

To apply this result to the chain $\tilde{Z}$, note that if $z = (w, x, y, u, \xi)$, then $\tilde{V}(z)^{1/2} = e^{\gamma w/2}$, so that $\tilde{\pi}$ possesses an exponential moment in w. Hence the stationary mean workload is also finite, implying that $\tilde{\pi} r < \infty$, since $r(z) \le w$. Finally then, we may conclude that $\beta_3(n) \to \beta_3$ a.s. as $n \to \infty$, thus defining β_3 and proving that $\beta_3(n)$ is consistent.

We summarize the above discussion with the following proposition.

Proposition 10 *For $i = 1, 2, 3$, the performance measures β_i, are well-defined, and the estimators $\beta_i(n)$ are strongly consistent (as $n \to \infty$).*

So we now turn to the error in the estimators. As before, this can be assessed through confidence intervals that derive from a central limit theorem. Again, in order for simplicity, we state the Markov chain central limit theorem under stronger conditions than are strictly necessary.

For a function $f : S \to \mathbb{R}$ with $\pi|f| < \infty$, let $\bar{f}(\cdot) = f(\cdot) - \pi f$. Also, let E_π denote the expectation operator over the path space of a Markov chain under initial distribution π.

Theorem 11 (MCCLT) *Suppose that the chain Φ satisfies the Lyapunov condition and is aperiodic. Then, for any function $f : \mathcal{S} \to \mathbb{R}$ with $f(z)^2 \le V(z)$ for all z,*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=0}^{n-1} f(\Phi_i) - \pi f\right) \Rightarrow \sigma N(0, 1),$$

where π is the stationary probability distribution of Φ, and

$$\sigma^2 = E_\pi[\bar{f}(\Phi_0)^2] + 2\sum_{k=1}^{\infty} E_\pi[\bar{f}(\Phi_0)\bar{f}(\Phi_k)]. \tag{10}$$

For a proof, see Theorem 17.0.1 of Meyn and Tweedie (1993).

We immediately obtain the following result.

Proposition 12 *We have that*

$$\sqrt{n}(\beta_i(n) - \beta_i) \Rightarrow \sigma_i N(0, 1)$$

as $n \to \infty$, for appropriately defined σ_i^2.

Thus, we see that just as in the terminating simulation case, the error in the estimator $\beta_i(n)$ is approximately normally distributed with mean 0 and variance σ_i^2/n.

This result serves as a foundation for constructing confidence intervals for β_i. One approach is to estimate σ_i^2 directly using the regenerative method, which is certainly easily applied to our example. But the method of batch means is, at least currently, more widely applicable, and the preferred method in commercial simulation software, and so we instead consider this approach.

Suppose that we have a sample path $\Phi_0, \Phi_1, \ldots, \Phi_{n-1}$. Divide this sample path into m batches of size b, where for convenience we assume that $n = mb$, so that the kth batch consists of observations $\Phi_{(k-1)b}, \ldots, \Phi_{kb-1}$. Now, for $k = 1, \ldots, m$, let M_k be the sample mean over the kth batch, i.e.,

$$M_k = \frac{1}{b}\sum_{i=(k-1)b}^{kb-1} f(\Phi_i),$$

and let $\bar{M}_m$ denote the sample mean of the m batch means $M_1, \ldots, M_m$. Finally, let

$$s_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(M_k - \bar{M}_m)^2$$

denote the sample variance of the M_k's. The method of batch means provides a confidence interval for πf of the form $\bar{M}_m \pm t s_m/\sqrt{m}$, for some constant t, and relies on the assumption that for large n, $(\bar{M}_m - \pi f)/(s_m/\sqrt{m})$ is approximately t-distributed, with $m-1$ degrees of freedom.

The MCCLT above implies that as $n \to \infty$ with m, the number of batches, held fixed, all of the batch means are asymptotically normally distributed with mean πf, and variance $m\sigma^2/n$. If each of the batch means are also asymptotically independent, then a standard result (see p. 173) of Rice (1988) for example) shows that the above confidence interval methodology is valid.

But how can we be sure that this asymptotic independence of the batch means will hold? A sufficient condition that supplies both the asymptotic independence, together with asymptotic normality, is that the chain Φ satisfy a functional central limit theorem; see Glynn and Iglehart (1990), from which much of the following discussion is adapted.

Definition 1 *Let Φ be a Markov chain on state space $\mathcal{S}$, and let $f : \mathcal{S} \to \mathbb{R}$. Define the continuous time process $Y = (Y(t) : t \geq 0)$ by $Y(t) = \Phi_{\lfloor t \rfloor}$. For $0 \leq t \leq 1$, let*

$$\bar{Y}_n(t) = n^{-1} \int_0^{nt} f(Y(s))\, ds$$

and set

$$\zeta_n(t) = n^{1/2}(\bar{Y}_n(t) - \kappa t),$$

for some constant κ. We say that Φ satisfies a functional central limit theorem (FCLT) if there exists an $\eta > 0$ such that $\zeta_n \Rightarrow \eta B$ as $n \to \infty$, where B denotes a standard Brownian motion.

Observe that if Φ satisfies a FCLT, then the jth batch mean M_j can be expressed as

$$\begin{aligned} M_j &= m[\bar{Y}_n(j/m) - \bar{Y}_n((j-1)/m)] \\ &= \kappa + n^{-1/2} m(\zeta_n(j/m) - \zeta_n((j-1)/m). \end{aligned}$$

Since the increments of Brownian motion are normally distributed, the FCLT then implies that the M_j's are asymptotically normally distributed with mean κ and variance $m\eta^2/n$, which is a conclusion that we had already reached. But the increments of Brownian motion are also independent, which implies that the M_j's are asymptotically independent, and this is the final result needed to ensure that the batch means confidence methodology outlined above is asymptotically valid.

So when can we be sure that Φ satisfies a FCLT? One sufficient condition is the following result.

Theorem 13 *Suppose that Φ satisfies the Lyapunov condition, and f is such that $f(z)^2 \leq V(z)$ for all z. If the constant σ^2 defined in (10) above is positive, then Φ satisfies a functional central limit theorem with $\kappa = \pi f$, and $\eta^2 = \sigma^2$.*

For a proof, see Theorems 17.4.4 and 17.5.3 of Meyn and Tweedie (1993).

Notice that we have already established that the conditions of Theorem 13 hold for all of our estimators $\beta_i(n)$. Thus, we immediately arrive at the conclusion that the method of batch means will yield asymptotically valid confidence intervals for each of the performance measures β_1, β_2 and β_3.

As in the terminating simulation case, the performance of these confidence interval procedures for finite n may be negatively impacted by bias. Of course, the bias depends on the initial distribution μ say of the chain Φ. Let E_μ denote the expectation operator over the path space of the chain Φ under initial distribution μ. Then the bias in the estimator $\beta_i(n)$ is given by $E_\mu \beta_i(n) - \beta_i$, for $i = 1, 2, 3$.

Let us first focus attention on $\beta_1(n)$. Let $f(w, x, y) = 1 - e^{-\lambda w}$. Borrowing a technique from Glynn (1995), we see that the bias in $\beta_1(n)$ under initial distribution μ is

$$\begin{aligned} & E_\mu \frac{1}{n} \sum_{i=0}^{n-1} (f(Z_i) - \pi f) \\ = \;& \frac{1}{n} E_\mu \sum_{i=0}^{\infty} (f(Z_i) - \pi f) - \frac{1}{n} E_\mu \sum_{i=n}^{\infty} (f(Z_i) - \pi f) \\ = \;& \frac{c}{n} + o(n^{-1}) \end{aligned}$$

provided that

$$c = E_\mu \sum_{i=0}^{\infty} (f(Z_i) - \pi f) < \infty. \tag{11}$$

So the bias in the estimator $\beta_1(n)$ will be of the order n^{-1} if (11) holds. This result holds in great generality. We in fact have the following result.

Theorem 14 *Suppose that Φ satisfies the Lyapunov condition and is aperiodic. Let π be the stationary probability distribution of Φ. If $f(z)^2 \leq V(z)$ for all z, and $\mu V < \infty$, then*

$$c = E_\mu \sum_{i=0}^{\infty} (f(\Phi_i) - \pi f) < \infty,$$

and so

$$E_\mu \frac{1}{n} \sum_{i=0}^{n-1} f(\Phi_i) - \pi f = \frac{c}{n} + O(q^n),$$

as $n \to \infty$, where $q < 1$.

The proof of this result is a straightforward extension of Theorem 16.0.1 of Meyn and Tweedie (1993).

We can conclude from this result that if the initial conditions are chosen appropriately (e.g., if Z_0 and $\tilde{Z}_0$

are chosen to be deterministic), then the bias of our three estimators is of the order n^{-1}.

Since the width of the batch mean confidence intervals is of the order $n^{-1/2}$, and the bias in the estimators is of the order n^{-1}, it follows that bias will typically only be an important factor for small sample sizes.

REFERENCES

Billingsley, P. 1986. *Probability and measure*. 2d ed. New York: Wiley.

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A guide to simulation*. 2d ed. New York: Springer-Verlag.

Glasserman, P. 1993. Filtered Monte Carlo. *Mathematics of Operations Research* 18:610–634.

Glynn, P. W. and P. Heidelberger. 1990. Bias properties of budget constrained simulations. *Operations Research* 38:801–814.

Glynn, P. W. and D. L. Iglehart. 1990. Simulation output analysis using standardized time series. *Mathematics of Operations Research* 15:1–16.

Glynn, P. W. 1995. Some new results on the initial transient problem. In *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. Lilegdon, and D. Goldsman, 165–170. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Henderson, S. G. and P. W. Glynn. 1999. Computing densities for Markov chains via simulation. Submitted for publication.

Law, A. M. and W. D. Kelton. 2000. *Simulation modeling and analysis*. 3d ed. New York: McGraw-Hill.

Meyn, S. P. and R. L. Tweedie. 1993. *Markov chains and stochastic stability.* New York: Springer-Verlag.

Rice, J. A. 1988. *Mathematical statistics and data analysis.* Pacific Grove, California: Wadsworth and Brooks/Cole.

Serfling, R. J. 1980 *Approximation theorems of mathematical statistics.* New York: Wiley.

AUTHOR BIOGRAPHY

SHANE G. HENDERSON is an assistant professor in the Department of Industrial and Operations Engineering at the University of Michigan. He was a lecturer in the Department of Engineering Science at the University of Auckland from 1997 to 1999. His research interests include discrete-event simulation, queueing theory and scheduling problems.