Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

VARIANCE REDUCTION IN THE SIMULATION OF CALL CENTERS

Pierre L'Ecuyer
Eric Buist

Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, C.P. 6128, Succ. Centre-Ville
Montréal (Québec), H3C 3J7, CANADA

ABSTRACT

We show via concrete illustrations how the variance can be reduced in the simulation of a telephone call center to estimate the fraction of calls answered within a given time limit. We examine the combination of a control variate and stratification with respect to a continuous input variable, and find that combining them requires care, because the optimal control variate coefficient is a function of the variable on which we stratify. In a setting where we compare two similar configurations of the center, we examine the combination of stratification with common random numbers. We show that proper use of common random numbers reduces the convergence rate of the variance of the difference of performance measures across the two systems.

1 INTRODUCTION

Telephone call centers, and more generally *contact centers* where mail, fax, e-mail, and Internet contacts are handled in addition to telephone calls, are important components of large organizations (Gans, Koole, and Mandelbaum 2003). One of the main problems in managing these centers is to optimize the number of *agents* who talk with customers over the phone, and the working schedules of these agents, under constraints on the quality of service and on admissible schedules. Large call centers are complex stochastic systems that can be analyzed realistically only by simulation; tractable queueing models oversimplify reality and are not very reliable. When simulation is combined with an optimization algorithm, its efficiency is a key issue, because optimization often requires thousands of simulation runs at different parameter settings (Ceşik and L'Ecuyer 2006). In this paper, we take a simulation model of a (simplified) call center inspired from real life and we experiment with different ways of reducing the variance without significantly increasing the computing cost.

The basic model is described in Section 2. In Section 3, we study the combination of stratification with a control variate (CV) for this model. The stratification is with respect to a uniform random number that drives the simulation. The optimal CV coefficient depends on the realization of this uniform and there is also an interdependence between the CV coefficients and the optimal allocation in strata. This type of combination is non-standard and requires some care. In Section 4, we study the efficiency of using common random numbers to estimate the sensitivity with respect to a parameter of the service time distribution. This is nontrivial mainly because the sample performance is discontinuous with respect to this parameter. Most examples are adapted from the future book of L'Ecuyer (2006). The simulations were all made with *ContactCenters*, a specialized simulation tool for contact centers (Buist and L'Ecuyer 2005) developed in Java with the SSJ library (L'Ecuyer and Buist 2005).

2 A SMALL MODEL OF A CALL CENTER

We consider a (simple) model of a telephone call center where agents answer incoming calls. Real-life call centers often have separate groups of agents having different combinations of *skills* that enable them to handle only a subset of the different types of calls. Here, to simplify the presentation, we assume a single type of agent and a single type of call. Otherwise, the model is strongly inspired by a real-life center in Canada. The techniques that we discuss also apply to more complex centers and other similar types of queueing systems.

Each day, the center operates for m hours. The number of agents answering calls and the arrival rate of calls vary during the day; we shall assume that they are constant within each hour of operation but depend on the hour. Let n_j be the number of agents in the center during hour j, for $j = 0,\ldots,m-1$. For example, if the center operates from 8 AM to 9 PM, then $m = 13$ and hour j starts at $(j+8)$ o'clock. All agents are assumed identical. If more than n_{j+1} agents are busy at the end of hour j, calls in progress are completed but new calls are answered only when there are fewer than n_{j+1} agents busy. After the center closes, ongoing calls are completed and calls already in the queue are answered, but no additional incoming call is taken.

The calls arrive according to a Poisson process with piecewise constant rate, equal to $R_j = B\lambda_j$ during hour j, where the λ_j are constants and B is a random variable with mean 1 that represents the *busyness factor* of the day. We suppose that B has the *gamma* distribution with parameters (α_0, α_0), i.e., with mean $\mathbb{E}[B] = 1$ and $\mathrm{Var}[B] = 1/\alpha_0$. The Poisson process assumption means that conditional on B, the number of incoming calls during any subinterval $(t_1, t_2]$ of hour j is a Poisson random variable with mean $(t_2 - t_1)R_j$ and that the arrival counts in any disjoint time intervals are independent. This type of arrival process model is motivated and studied by Whitt (1999) and Avramidis, Deslauriers, and L'Ecuyer (2004).

Incoming calls form a FIFO queue for the agents. A call *abandons* (and is lost) when its waiting time exceeds its *patience time*. The patience times of calls are assumed to be i.i.d. random variables with the following distribution: with probability p the patience time is 0 (so the person hangs up if no agent is available immediately), and with probability $1-p$ it is exponential with mean $1/\nu$. The *service times* are i.i.d. *gamma* random variables with parameters (α, γ), i.e., with mean α/γ and variance α/γ^2.

We want to estimate $g(s_0)$, the fraction of calls whose waiting time is less than s_0 seconds (including those who abandoned before s_0 seconds), for a given threshold s_0, over infinite time horizon (i.e., an infinite number of days). This $g(s_0)$ is called the *service level* and is by far the most widely used measure of quality of service in call centers. In many cases, it is even regulated by law. Let A be the number of arriving calls during the day and $X = G(s_0)$ the number of those calls waiting less than s_0 seconds. The expected number of arrivals is $a = \mathbb{E}[A] = \sum_{j=0}^{m-1} \lambda_j$ and we have $g(s_0) = \mathbb{E}[G(s_0)]/a$. Since a is known, here we will estimate $\mu = a \cdot g(s_0) = \mathbb{E}[G(s_0)]$.

We simulate the model for n days. For each day i, let A_i be the number of arrivals and $X_i = G_i(s_0)$ the number of calls who waited less than s_0 seconds. A straightforward (or crude) unbiased estimator of μ is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

with variance $\mathrm{Var}[\bar{X}_n] = \mathrm{Var}[X_i]/n$. We can estimate $\mathrm{Var}[X_i]$ by the empirical variance and a confidence interval can be computed as usual, using the normal approximation.

For our numerical illustration, we use the following parameters, where the time is measured in seconds: $\alpha_0 = 10$, $p = 0.1$, $\nu = 0.001$, $\alpha = 1.0$, $\gamma = 0.01$ (so the mean service time is 100 seconds), and $s_0 = 20$. The center starts empty and operates for 13 one-hour periods. The number of agents and the arrival rate in each period are given in Table 1.

We ran a simulation experiment with $n = 10000$. The sample mean and sample variance of the X_i's were $\bar{X}_n = 1418.0$ and $S_n^2 = 77418$, respectively. The estimated variance of $\bar{X}_n$ is thus 7.74. Next, we show how to reduce this variance.

3 COMBINING STRATIFICATION WITH A CONTROL VARIATE

3.1 Stratification on B

Noting that the value of B is obviously an important source of variance for $X = G(s_0)$, a first idea would be to stratify on B. We can partition the set of all possible values of B in k *strata* with an average of $m = n/k$ observations per strata. Assume that B is generated by inversion from a single $\mathrm{U}(0,1)$ random variate U, that is, $B = F_{\mathrm{B}}^{-1}(U)$ where F_{B} is the distribution function of the busyness factor and U is uniformly distributed over the interval $(0,1)$. We stratify on U instead of B. For this, we simply partition $(0,1)$ into k subintervals of length $1/k$. These intervals determine k strata. To generate an observation (i.e., simulate one day) in stratum s, for $s = 1, \ldots, k$, we simply generate U uniformly in $[(s-1)/k, s/k)$ and put $B = F_{\mathrm{B}}^{-1}(U)$. Suppose we generate n_s observations in stratum s for each s, where the n_s's are positive integers such that $n = n_1 + \cdots + n_k$. Let $X_{s,1}, \ldots, X_{s,n_s}$ denote the n_s i.i.d. observations X in stratum s. The (unbiased) *stratified estimator* of μ in this case is (Cochran 1977):

$$\bar{X}_{\mathrm{s},n} = \frac{1}{k}\sum_{s=1}^{k} \hat{\mu}_s \qquad \text{where} \qquad \hat{\mu}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} X_{s,i} \tag{1}$$

is the sample mean within stratum s. Let $\sigma_s^2 = \mathrm{Var}[X \mid S = s]$, the conditional variance of X given that we are in stratum s. Then,

$$\mathrm{Var}[\bar{X}_{\mathrm{s},n}] = \frac{1}{k^2}\sum_{s=1}^{k} \sigma_s^2/n_s \tag{2}$$

and an unbiased estimator of this variance is

$$S_{\mathrm{s},n}^2 = \frac{1}{k^2}\sum_{s=1}^{k} \hat{\sigma}_s^2/n_s, \tag{3}$$

where $\hat{\sigma}_s^2$ is the sample variance of $X_{s,1}, \ldots, X_{s,n_s}$, assuming that $n_s \geq 2$.

Stratification with *proportional allocation* takes $n_s = n/k$ for all s. Then, (2) simplifies to

$$\mathrm{Var}[\bar{X}_{\mathrm{sp},n}] = \frac{1}{nk}\sum_{s=1}^{k} \sigma_s^2 \tag{4}$$

where $\bar{X}_{\mathrm{sp},n}$ denotes the corresponding version of (1). The *optimal allocation*, which minimizes the variance (2) with respect to $n_1, \ldots, n_k$ under the constraints that $n_s > 0$ for each s and $n_1 + \cdots + n_k = n$ for a given n, is easily found by using a Lagrange multiplier; we must take n_s proportional to $p_s\sigma_s$: $n_s^* = n\sigma_s/\bar{\sigma}k$ where $\bar{\sigma} = \sum_{s=1}^{k} \sigma_s/k$. (We neglect the

Table 1: Number of Agents n_j and Arrival Rate λ_j (per hour) for 13 one-hour Periods in the Call Center

j	0	1	2	3	4	5	6	7	8	9	10	11	12
n_j	4	6	8	8	8	7	8	8	6	6	4	4	4
λ_j	100	150	150	180	200	150	150	150	120	100	80	70	60

rounding of n_s^* to an integer and assume that $n_s \geq 2$.) If $\bar{X}_{\mathrm{so},n}$ denotes the estimator with optimal allocation, we have $\mathrm{Var}[\bar{X}_{\mathrm{so},n}] = \bar{\sigma}^2/n$. Putting these pieces together, the variance can be decomposed as follows (Cochran 1977):

$$\begin{aligned} & \mathrm{Var}[\bar{X}_n] \\ = \; & \mathrm{Var}[\bar{X}_{\mathrm{sp},n}] + \frac{1}{nk}\sum_{s=1}^{k}(\mu_s-\mu)^2 \\ = \; & \mathrm{Var}[\bar{X}_{\mathrm{so},n}] + \frac{1}{nk}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2 + \frac{1}{nk}\sum_{s=1}^{k}(\mu_s-\mu)^2. \end{aligned}$$

The first sum in the last line represents the variability due to the different standard deviations among strata and the second sum represents the variability due to the differences between stratum means. Proportional allocation eliminates the last sum while optimal allocation also eliminates the first.

Note that for a given n, we could take a larger k together with a smaller $m = n/k$, or a smaller k together with a larger m. (We assume that both m and k are positive integers.) A larger k gives more variance reduction, because there are more strata. But the marginal gain converges to zero when k increases, because when the number of strata becomes very large, the variation of B becomes a negligible source of variance relative to the overall variance. On the other hand, with a larger m, we have a more accurate estimator of the variance of the stratified estimator.

3.2 Combining with a Control Variate

Control variates (CVs) for simulation are studied, e.g., by Lavenberg and Welch (1981) and Glynn and Szechtman (2002). The way we combine a CV with stratification here seems novel. A simple CV that quickly comes to mind here is A, the number of arrivals in the day. A standard (naive) way of using this CV (without the stratification) is by subtracting $A - \mathbb{E}[A] = A - a$ multiplied by a *constant* coefficient β:

$$X_{\mathrm{c}} = X - \beta[A-a].$$

The optimal coefficient β is $\beta^* = \mathrm{Cov}[A,X]/\mathrm{Var}[A]$. We tried this CV estimator with $n = 10^4$ simulation runs, estimating β^* by computing the empirical covariance and variance from the same runs (this gives a slightly biased estimator, but the bias is negligible here with this large n). We obtained a mean of 1418 with a variance of 36230 (empirical). This improves over the original variance of 77418 by a factor of 2.13.

To use the control variate jointly with stratification as we propose in this paper, we must be careful, because the expectation of A_i depends on B so it changes between strata, and the optimal CV coefficient also differs across strata. Let $C = A - \mathbb{E}[A \mid B]$ be the centered CV. The CV estimator conditional on $U = u$ is then

$$X_{\mathrm{c}}(u) = X - \beta(u)[A - \mathbb{E}[A \mid U = u]].$$

The optimal CV coefficient $\beta(u)$ as a function of $U = F_{\mathrm{B}}(B) = u$, is

$$\beta^*(u) = \frac{\mathrm{Cov}[C,X|U=u]}{\mathrm{Var}[C|U=u]} = \frac{E[C\cdot X \mid U=u]}{E[C^2 \mid U=u]}. \qquad (5)$$

This function can be approximated by approximating the two functions $q_1(u) = E[C\cdot X \mid U=u]$ and $q_2(u) = E[C^2 \mid U=u]$. These functions can be estimated from a sample $\{(U_i,C_i,X_i), i=1,\ldots,n\}$ of n realizations of (U,C,X), e.g., by fitting a curve $\hat{q}_1$ to the points (U_i,C_iX_i) and another curve $\hat{q}_2$ to the points (U_i,C_i^2). Note that a β that depends on u can do much better than a fixed β (the same for all values of u) when $\beta^*(u)$ is far from being a constant.

The variance of the controlled estimator, conditional on $U=u$, is $\sigma_{\mathrm{c}}^2(u) = \mathrm{Var}[X_{\mathrm{c}}(u)] = \mathrm{Var}[X - \beta^*(u)C \mid U=u]$. The variance in stratum s is

$$\begin{aligned} \sigma_s^2 &= \mathrm{Var}[X_{\mathrm{c}}(U^{(s)})] \\ &= \mathbb{E}[\mathrm{Var}[X_{\mathrm{c}}(U^{(s)})]] + \mathrm{Var}[\mathbb{E}[X_{\mathrm{c}}(U^{(s)})]] \\ &= \int_{(s-1)/k}^{s/k} \mathrm{Var}[X - \beta^*(u)C \mid U=u]du \end{aligned}$$

where $U^{(s)}$ is uniformly distributed over $[(s-1)/k, s/k)$. The optimal allocation takes n_s proportional to this σ_s.

To approximate β^* and σ_{c}^2 as functions of u, a simple idea is to use (rather crude) quadratic approximations of the form $\beta(u) = \beta_0 + \beta_1 u + \beta_2 u^2$ and $\tilde{\sigma}_{\mathrm{c}}(u) = \sigma_0 + \sigma_1 u + \sigma_2 u^2$, as follows. We first estimate $\beta^*(u)$ and $\sigma_{\mathrm{c}}^2(u)$ at $u = u_1 = 0.2$, $u_2 = 0.5$ and $u_3 = 0.8$, from n_0 (pilot) simulation runs at each of those values of u (i.e., with B fixed at $F_{\mathrm{B}}^{-1}(u_j)$). Let b_j and v_j denote the corresponding estimates of $\beta^*(u_j)$ and $\sigma_{\mathrm{c}}(u_j)$, respectively. We can then compute $(\beta_0,\beta_1,\beta_2)$ so that the quadratic function interpolates the three points (u_j,b_j). Similarly, we compute $(\sigma_0,\sigma_1,\sigma_2)$ for which the function $\tilde{\sigma}_{\mathrm{c}}$ interpolates the points (u_j,v_j). We then standardize this function $\tilde{\sigma}_{\mathrm{c}}(u)$, by dividing it by

$$\bar{\sigma}_{\mathrm{c}} = \int_0^1 \tilde{\sigma}_{\mathrm{c}}(u)du = \sigma_0 + \sigma_1/2 + \sigma_2/3,$$

Table 2: Interpolations and Least Squares Quadratic Approximations Obtained for $\beta^*(u)$, $\sigma(u)$, and $\sigma_c(u)$

$\beta^*(u)$, 3 pts	$0.9+0.271u-1.44u^2$
$\sigma_c(u)$, CV	$8.457-0.69u+67.894u^2$
$\sigma(u)$, no CV	$37.644-44.806u+78.582u^2$
$\beta^*(u)$, 50 pts	$0.921+0.197u-1.527u^2$
σ_c CV, 50 pts	$8.47-28.164u+113.89u^2$
σ no CV, 50 pts	$35.023-67.468u+128.637u^2$

so that it gives a quadratic approximation of $\sigma_c(u)/\int_0^1 \sigma_c(u)du$. The optimal allocation to any stratum is easy to obtain after this standardization: it is the average of the standardized function over the stratum, multiplied by $m = n/k$.

In case where we use stratification *without* the CV, the optimal allocation to strata can be estimated in the same way, except that there is no $\beta(u)$. The function $\sigma_c^2(u)$ is replaced by $\sigma^2(u) = \mathrm{Var}[X \mid U = u]$. Note that this function $\sigma(u)$ differs from $\sigma_c(u)$, so it must be approximated separately and the optimal allocation is different.

Based on pilot runs with $n_0 = 200$ per testing point, we obtained the quadratic approximations for β, σ_c, σ given in Table 2. We ran a simulation experiment with $n = 10^4$ and $k = 100$ strata. Our estimates of the terms $\mathrm{Var}[\bar{X}_n]$, $\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$, $(1/nk)\sum_{s=1}^k(\sigma_s-\bar{\sigma})^2$, and $(1/nk)\sum_{s=1}^k(\mu_s-\mu)^2$ in the variance decomposition (the first term is the sum of the three other) are given in Table 3. In comparison with standard Monte Carlo, the variance is reduced by a factor of

28.1 for proportional allocation without CV;
35.4 for with optimal allocation without CV;
35.8 for with proportional allocation with CV;
60.4 for with optimal allocation with CV.

Another (simpler) option that bypasses all functional estimations is to use stratification with proportional allocation, combined with the CV with a *constant* coefficient β. For this case, we find that the optimal β is approximately 0.45. This is labeled "CV, β const" in the table. The variance is reduced by a factor of 30.6.

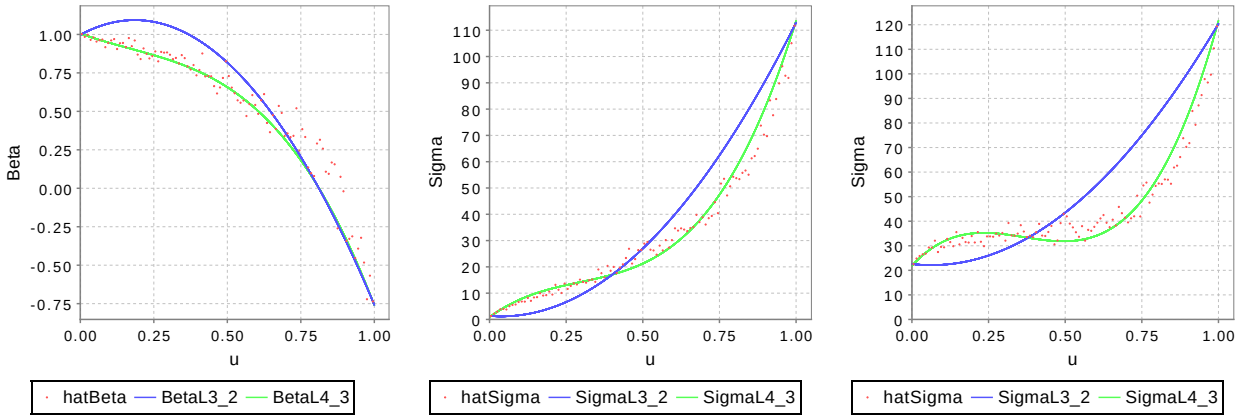
We made further experiments to see if we could improve the quality of approximation of the functions $\beta(u)$, $\sigma_c(u)$, and $\sigma(u)$. We tried the following alternatives: (a) we fitted an interpolating quadratic polynomial to three points, at 0.005, 0.5, 0.995; (b) we fitted an interpolating cubic polynomial to four points, at 0.005, 0.335, 0.665, 0.995; (c) instead of interpolating, we used least squares to fit a quadratic curve on 100 points at $u_i = (i+0.5)/100$, for $i = 0,\ldots,99$ (the function values were estimated from 100 simulation runs at each of these points); (d) we fitted a cubic polynomial by least squares to the same 100 points; (e) we fitted a cubic spline to these 100 points; (f) we fitted a smoothing cubic spline to 50 points. With the latter method, a constraint is imposed on the root mean square error at the selected points (this is called the *smoothing factor*) and the algorithm find the smoothest possible spline that satisfies the constraint.

Figure 1 shows the quadratic and cubic interpolations (a) and (b), for the three functions, with black points representing results of pilot runs. Figure 2 gives the quadratic and cubic polynomial least-squares fits (c) and (d). The cubic polynomial least-squares fit (d) and the splines (e) and (f) (not shown due to lack of space) match the points a bit better and are about equally good between themselves. The plots give an idea of how $\beta^*(u)$, $\sigma_c(u)$, and $\sigma(u)$ behave as functions of u. The optimal CV coefficient $\beta^*(u)$, which has the same sign as $\mathrm{Cov}[C,X|U=u]$, is positive for small u and negative for large u. This can be explained intuitively as follows: when u is small, the load on the system is small and the agents are not very busy, so a small increase in the number of arrivals tends to increase $G(s_0)$. If u is large, on the other hand, the agents are occupied most of the time, so a few more arrivals increases the waiting time of several calls and tends to *decrease* the number of calls answered with s_0 seconds. The functions σ_c and σ, on the other hand, increase sharply when u approaches 1, which means that the estimators have much more variance when the arrival rate is very large (the system is highly loaded).

Table 3 presents the results of a simulation experiment with $n = 10^4$ replications divided into 100 strata. We compare the empirical variances with the following three methods for estimating the functions: a quadratic fit with three points, a quadratic least squares approximation with 50 points, and a cubic smoothing spline with 50 points. The smoothing splines are not really doing better for this example and they require more overhead to estimate the function with pilot runs. Selection of the smoothing factor is also not automatic.

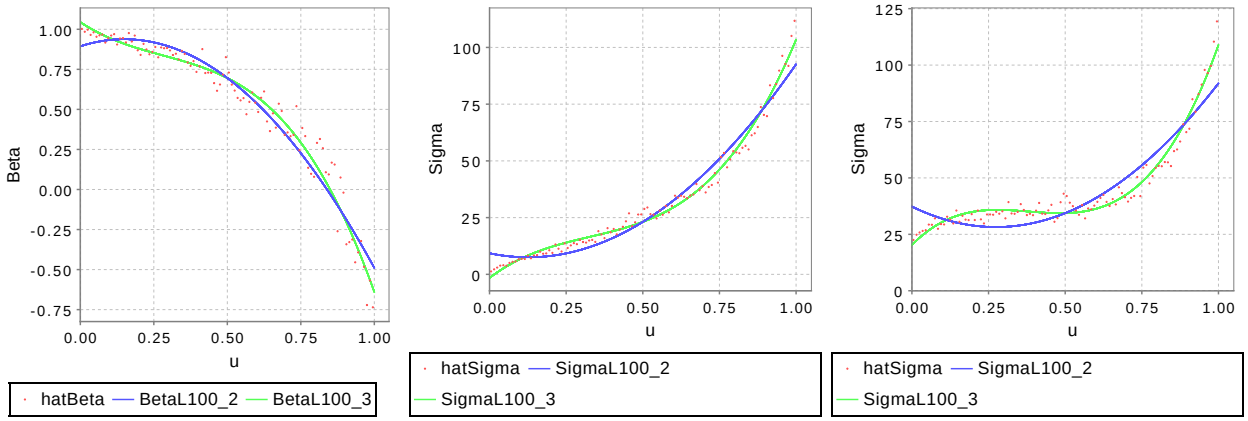
4 COMPARING SYSTEMS WITH COMMON RANDOM NUMBERS

We now examine a situation where we want to compare two very similar configurations of the call center. This is often required in optimization settings and for sensitivity analysis. Configuration 1 (the base configuration) is the same as in the previous section. In Configuration 2, we may have a slightly different number of agents in one or more periods, or slightly different parameter values for one of the distributions (e.g., service times, patience times, arrival process). Here, we consider a small change in a parameter of the service time distribution. We have selected a continuous parameter, because we want to analyze what happens when the size of the change converges to zero.



(i) $\beta^*(u)$ (ii) $\sigma_c(u)$ (iii) $\sigma(u)$

Figure 1: The Functions $\beta^*(u)$, $\sigma_c(u)$, and $\sigma(u)$ Approximated by Polynomial Interpolation



(i) $\beta^*(u)$ (ii) $\sigma_c(u)$ (iii) $\sigma(u)$

Figure 2: The Functions $\beta^*(u)$, $\sigma_c(u)$, and $\sigma(u)$ Approximated by Least Squares on 100 Points

4.1 Sensitivity to the Mean Service Time

The service times in our model are i.i.d. gamma(α,γ), with mean α/γ and variance α/γ^2. In the base configuration, we have $\alpha = 1$ and $\gamma = \gamma_1 = 1/100$. (With $\alpha = 1$, the service times are exponential.) We want to study the effect of changing slightly the second parameter to $\gamma = \gamma_2 = \gamma_1/(1-\delta)$ for a small δ. The effect of this is to multiply the mean service time by $1-\delta$ while keeping the coefficient of variation unchanged. The distribution of $X = G(s_0)$ and its expectation, $\mu(\gamma) = \mathbb{E}_\gamma[X]$, now depend on γ. Suppose we want to estimate $\mu(\gamma_2) - \mu(\gamma_1)$.

We simulate n days at each server speed. Let $X_{1,i}$ and $X_{2,i}$ be the realizations of $G(s_0)$ on day i, with $\gamma = \gamma_1$ and $\gamma = \gamma_2$, respectively. Define $\Delta_i = X_{2,i} - X_{1,i}$ and let $\bar{\Delta}_n = (1/n)\sum_{i=1}^n \Delta_i$ be the (crude) estimator of the difference $\mu(\gamma_2) - \mu(\gamma_1)$. To compute $X_{1,i}$ and $X_{2,i}$, we can use either (i) independent random numbers (IRNs) or (ii) common random numbers (CRNs), i.e., the same underlying uniform random numbers for the two values of γ (Bratley, Fox, and Schrage 1987, L'Ecuyer and Perron 1994). We have

$$\mathrm{Var}[\Delta_i] = \mathrm{Var}[X_{1,i}] + \mathrm{Var}[X_{2,i}] - 2\mathrm{Cov}[X_{1,i}, X_{2,i}].$$

With IRNs, the covariance term is zero. The aim of CRNs is to make this covariance positive. By using the same random numbers at the same places for both systems as far as this is achievable, the responses $X_{1,i}$ and $X_{2,i}$ are expected (intuitively, at least) to be strongly positively correlated, especially if γ_1 and γ_2 are close.

To see how the same random numbers can be used at approximately the same places in this example, we distinguish four types of random numbers: Those used to generate (1) the busyness factor in the morning; (2) the interarrival times; (3) the service times; and (4) the patience times. To make

Table 3: Stratification with a Control Variate: Variance Comparisons

	Quadratic interpolation			Least squares, 50 points		Cubic spline	
	No CV	CV, β const.	CV, $\beta(u)$	No CV	CV, $\beta(u)$	No CV	CV, $\beta(u)$
$\bar{X}_n$	1419.48	1419.30	1418.59	1419.29	1418.88	1418.64	1419.08
$n\mathrm{Var}[\bar{X}_n]$	77083	77158	76713	77556	76529	77312	76578
$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	2743	2515	2142	2720	2166	2754	2098
$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	2178	—	1269	2193	1294	2181	1234
$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	565	—	873	527	872	573	864
$\frac{1}{k}\sum_{s=1}^{k}(\mu_s-\mu)^2$	74341	74643	74571	74836	74363	74558	74481

sure that the same sequences of random numbers are employed within each category, we attach a different random number stream to each category. We initialize these four generators to the same seeds for both server speeds and we generate *all random variates by inversion*.

Do we really need to use inversion to maintain the synchronization for the CRNs? For this particular example, the answer is *no*. The only place where this could make a difference is the generation of service times, because for the other random variates, nothing is changed across the two configurations. For a general gamma distribution, inversion is quite time-consuming; much faster acceptance-rejection methods are available. One such method, implemented in SSJ, generates a gamma random variable of mean 1 with the required shape parameter α, then multiplies the value by the desired mean. In our example, since only the scale parameter is changed, the acceptance-rejection part of the algorithm does exactly the same thing for both configurations; only the multiplication by the desired mean changes. For this reason, acceptance-rejection here maintains synchronization exactly as inversion. To check this, we actually tried replacing inversion with acceptance-rejection in the experiments described in what follows, and we found no significant difference.

In our attempt to use the same random numbers at exactly the same places for both systems, we meet a difficulty with the synchronization: Due to the different service times, the abandonments will eventually not be the same for γ_1 and γ_2. A given call may abandon in one system and not in the other. For such a call, the service time must be generated in the first case but does not have to be generated in the second case. In this context, we can generate service times:

(a) for all calls (the abandoned calls will have unused service times), or

(b) only for the calls that are actually answered.

The rationale for (a) is to make sure that the same calls have the same service times in both configurations, regardless of abandonments. The argument for (b) is for the sequence of service times that are effectively employed to be the same for both systems, even though these service times may belong to different calls because of different abandonment decisions. So if a call has a very long service time for parameter γ_1, this very long service time will also appear (perhaps for another call) for γ_2 under (b). Under (a), this long service time could be unused for γ_2.

Likewise, the patience time can be generated:

(c) for all arriving calls, to maintain synchronization, or

(d) only for the calls whose service does not start immediately upon arrival.

By combining these choices, we obtain four different possibilities for the synchronization strategy, and it is unclear a priori which one is best.

Table 4 reports empirical results for $\delta = 0.1$, 0.01, and 0.001, with $n = 10^4$. This corresponds to a reduction of the mean service time by 10%, 1%, and 0.1%, respectively. The table gives the sample mean and the sample variance of the Δ_i's for simulations with independent random numbers (IRNs), simulations with CRNs without synchronization, and simulations with CRNs with the different types of synchronizations just described. For CRNs *without synchronization*, we just took all the random numbers needed in the simulation sequentially from a single stream instead of from four different streams, and we generated random variates only when they were needed (strategy (b + d)). Thus, for example, an inter-arrival time for one system configuration could be generated from the same random number as a service time for the other configuration. All the pairs $(\bar{\Delta}_n, \widehat{\mathrm{Var}}[\Delta_i])$ in the table were obtained by independent simulations.

With any of the four combinations of synchronization approaches, CRNs reduce the variance by a huge factor. The smaller δ is, the more the variance is reduced. With IRNs, we have $\mathrm{Var}[\Delta_i] = \mathrm{Var}[X_{1,i}] + \mathrm{Var}[X_{2,i}] \approx 2\,\mathrm{Var}[X_i]$, so the variance does not depend much on δ. (It is slightly larger for $\delta = 0.1$ because $\mathrm{Var}[X_{2,i}]$ is significantly larger in that case.) The difference between the four IRN results is just noise. With CRNs, the variance diminishes rapidly when $\delta \to 0$. A theoretical analysis of the convergence is provided in Section 4.2. When $\delta \to 0$, $\mathbb{E}[\Delta_i]$ becomes closer to 0 and thus harder to estimate. When δ is very small, the finite-difference estimators with IRNs are practically useless (too noisy), whereas those with CRNs remain viable.

Table 4: Effect of Change in the Mean Service Time for the Call Center, with CRNs, for $n = 10^4$

Method	$\delta = 0.1$		$\delta = 0.01$		$\delta = 0.001$	
	$\bar{\Delta}_n$	$\widehat{\text{Var}}[\Delta_i]$	$\bar{\Delta}_n$	$\widehat{\text{Var}}[\Delta_i]$	$\bar{\Delta}_n$	$\widehat{\text{Var}}[\Delta_i]$
IRN (a + c)	76.93	192680	7.48	160486	-0.135	157745
IRN (a + d)	76.57	192374	7.00	160168	-0.560	157592
IRN (b + c)	76.96	192034	7.16	159605	-0.216	156794
IRN (b + d)	77.14	192495	7.34	159970	-0.215	157136
CRN, no sync. (b + d)	76.08	11036	6.71	2417	0.203	1979
CRN (a + c)	78.75	9307	8.48	122	0.862	5
CRN (a + d)	78.78	9502	8.51	389	0.786	197
CRN (b + c)	78.73	9714	8.14	237	0.821	21
CRN (b + d)	78.90	9646	8.42	312	0.819	81

CRNs with (b + d) without synchronization reduces the variance much less than the good synchronization schemes, but nevertheless improves over IRNs by a significant factor. Understanding exactly why it works would require further investigation, but most of the explanation might be that (i) it takes some time before the synchronization is lost and (ii) the important variable B_i is always generated with the same random number, and thus takes a common value for both systems. With strategy (a + c), it turns out that for this particular example, synchronization is maintained across the two systems even with a single stream, because the two systems get a common value of B_i and the same sequence of arrival events, each call requires exactly three random numbers upon arrival, and no random numbers are needed anywhere else. For this reason, CRNs with a single stream give the same variance reduction as CRNs with the four different streams as described earlier.

Between the synchronization strategies for the CRNs, the (a + c) combination is the best performer, followed by (b + c). So it is better in this example to generate service and patience times for all calls, and simply discard the values that are not needed. The choice of strategy makes a significant difference when δ is very small, but not much when δ is larger (e.g., 0.1). We must underline that this observation should not be taken as a general rule: there are similar situations where the best synchronization strategy would be (b + d) instead. So in practice, it is worth trying and comparing.

To reduce the variance further, we can combine the use of CRNs with stratification and a CV, as in Section 3. For the CV, with CRNs, we use the number A of arrivals (not the difference in number of arrivals between the two configurations, because the number of arrivals is the same for both configurations). The functions β^*, σ_c, and σ are estimated in a similar way. Table 5 presents our results for this case, for $n = 10^4$ replications with 100 strata, with (a + c) to synchronize the random numbers. We used quadratic interpolation to estimate the functions. The empirical results indicate that combining CRNs with stratification provides an additional variance reduction by a significant factor. The CV, on the other hand, does not bring additional gain. For $\delta = 0.1$, the variance goes down approximately from 192000 to 9300 with the CRNs and to 230 when the three methods are combined. For $\delta = 0.001$, it goes from about 157000 to 5 with the CRNs (a + c) and to 2 with the stratification.

4.2 Bound on Convergence Rate When $\delta \to 0$

With CRNs, $X = X(\delta)$ is a function of δ. If $X(\delta)$ was a continuous function of δ with a denumerable number of non-differentiability points with probability 1, then $\text{Var}[(X(\delta) - X(0))/\delta]$ would be bounded by a constant when $\delta \to 0$ and we would have

$$\mathbb{E}\left[\lim_{\delta \to 0} \frac{X(\delta) - X(0)}{\delta}\right] = \left.\frac{\partial g(s_0)}{\partial \delta}\right|_{\delta=0}$$

where the pathwise limit inside the expectation is the *infinitesimal perturbation analysis* estimator of the derivative of $g(s_0)$ with respect to δ (L'Ecuyer 1990, Glasserman 1991, Glasserman and Yao 1992, L'Ecuyer and Perron 1994). In that case, $\text{Var}[X(\delta) - X(0)] = \mathcal{O}(\delta^2)$ when $\delta \to 0$.

Here the service times change continuously with δ, but $X(\delta)$ is a piecewise constant function of δ that takes only integer values, so it is definitely not continuous. The next proposition states that under the CRN strategy (a + c), $\text{Var}[X(\delta) - X(0)] = \mathcal{O}(\delta^{1-\varepsilon})$ for any $\varepsilon > 0$ when $\delta \to 0$. This is certainly not as good as $\mathcal{O}(\delta^2)$, but much better than the $\mathcal{O}(1)$ rate that we get with IRNs. Empirically, the rate seems even better than $\mathcal{O}(\delta)$: the variance for CRN with (a + c) in Table 4 is divided by much more than 10 when δ is divided by 10.

Proposition 1 *We have* $\text{Var}[\Delta] = \mathcal{O}(\delta^{1-\varepsilon})$ *where the hidden constant depends on* ε *but not on* δ.

The main ingredient in our proof of this proposition is a lemma that bounds the probability that $X(\delta) \neq X(0)$, as a

Table 5: Effect of Change in the Mean Service Time for the Call Center, with CRNs, Stratification, and CV, for $n = 10^4$ and $k = 100$

	$\delta = 0.1$			$\delta = 0.01$			$\delta = 0.001$		
	No CV	CV β const	CV $\beta(u)$	No CV	CV β const	CV $\beta(u)$	No CV	CV β const	CV $\beta(u)$
$\bar{\Delta}_n$, CRN	77.07	77.00	77.14	8.270	8.300	8.210	0.816	0.837	0.808
$\mathrm{Var}[\bar{\Delta}_n]$, CRN	9086	9061	9075	121	119	115	5.0	5.1	4.9
$\mathrm{Var}[\bar{\Delta}_{\mathrm{sp},n}]$, CRN	464	415	417	33	32	31	4.2	4.1	4.1
$\mathrm{Var}[\bar{\Delta}_{\mathrm{so},n}]$, CRN	267	—	230	18	—	18	2.0	—	1.9
$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s - \bar{\sigma})^2$, CRN	197	—	186	15	—	13	2.2	—	2.1
$\frac{1}{k}\sum_{s=1}^{k}(\mu_s - \mu)^2$, CRN	8622	8646	8658	88	87	84	0.8	1.0	0.8

function of δ. For this, we need some notation. In the original model ($\delta = 0$), the service time of the jth call (by order of arrival) is $S_j = G^{-1}(U_j)$ where G is the gamma(α,γ) distribution function and the U_j's are i.i.d. U$(0,1)$. Let T_j denote the arrival time of call j, P_j its patience time, and W_j the time at which it is answered. Multiplying the parameter γ by $1/(1-\delta)$ changes the service time from S_j to $S_j(\delta) = (1-\delta)S_j$, and the answering time from W_j to $W_j(\delta)$. Let $D_j = |W_j - W_j(\delta)|$. This call has *good service* in the original model if and only if $W_j \le V_j$, where $V_j = T_j + \min(P_j, s_0)$ is its *virtual threshold time* (VTT). Note that the V_j's are independent of the S_j's and of δ. Call j switches from bad to good service when γ_1 changes to $\gamma_1/(1-\delta)$ if and only if

$$W_j(\delta) \le V_j < W_j. \tag{6}$$

Similarly, call j switches from good to bad service if and only if

$$W_j \le V_j < W_j(\delta). \tag{7}$$

In general, the status of call j changes if and only if

$$\min\{W_j, W_j(\delta)\} \le V_j < \max\{W_j, W_j(\delta)\}. \tag{8}$$

Let $\bar{\lambda}(u)$ be the maximum arrival rate during the day conditional on $B = b = F_{\mathrm{B}}^{-1}(u)$. Let $\mathbb{P}_u$ and $\mathbb{E}_u$ denote the corresponding conditional probability and conditional expectation.

Lemma 1 *Conditional on $B = b = F_{\mathrm{B}}^{-1}(u)$, we have*

$$\begin{aligned} & \mathbb{P}_u[X(\delta) \neq X(0)] \\ \le\ & \mathbb{E}_u[A^2]\alpha\bar{\lambda}(u)\delta/\gamma + \mathbb{E}_u[A^3]\mathscr{O}(\delta^2) \\ =\ & \mathscr{O}(\delta b^3). \end{aligned}$$

Proof. Recall that the patience time is exponential with parameter ν. For a small $\varepsilon > 0$, a given time interval $[t, t+\varepsilon)$ can contain one (or more) of the V_j's only if a call arrives in the time interval $[t - s_0, t - s_0 + \varepsilon)$ and reaches its VTT before abandoning, or if a call arrives at time x for $t - s_0 \le x \le t + \varepsilon$ and abandons during the interval $[t, t+\varepsilon)$. The probability that one of these two events occurs is bounded by

$$\begin{aligned} & \bar{\lambda}(t - s_0)\,\varepsilon\, e^{-\nu s_0} + \int_{t-s_0}^{t} \bar{\lambda}(x)\,\varepsilon\nu e^{-\nu(t-x)}dx + o(\varepsilon) \\ \le\ & \varepsilon\,\bar{\lambda}(u) + o(\varepsilon) \end{aligned}$$

(where the function $\bar{\lambda}$ depends on u). This gives an upper bound on the probability that $[t, t+\varepsilon]$ contains V_j for a fixed j. By integrating this with respect to t over $[\min\{W_j, W_j(\delta)\}, \max\{W_j, W_j(\delta)\}]$, and taking $\varepsilon \to 0$, we find that the probability that (8) occurs cannot exceed $\bar{\lambda}(u)D_j$.

Let J^* be the smallest integer $j > 0$ for which (8) holds, i.e, the index of the first call that switches status. If there is none, put $J^* = \infty$. For $j \le J^*$, we have

$$D_j = |W_j - W_j(\delta)| \le \sum_{\ell=1}^{j-1}(S_\ell - S_\ell(\delta)) = \delta\sum_{\ell=1}^{j-1} S_\ell \stackrel{\mathrm{def}}{=} \xi_j.$$

Combining the last two bounds, we obtain that

$$\begin{aligned} & \mathbb{P}_u[J^* = j \mid W_j, W_j(\delta)] \\ \le\ & \mathbb{P}_u[\min\{W_j, W_j(\delta)\} \le V_j < \max\{W_j, W_j(\delta)\} \mid W_j, W_j(\delta)] \\ \le\ & \bar{\lambda}(u)D_j \\ \le\ & 1 - \exp[-\bar{\lambda}(u)D_j] \\ \le\ & 1 - \exp[-\bar{\lambda}(u)\xi_j] \\ =\ & 1 - \prod_{\ell=1}^{j-1}\exp[-\bar{\lambda}(u)\delta S_\ell]. \end{aligned}$$

Observe that the ξ_j's are independent of the V_j's, because the service times are independent of the arrival process. Recall also that S_ℓ has moment generating function $M(t) = \mathbb{E}[\exp[tS_\ell]] = (1 + t/\gamma)^{-\alpha} = 1 - \alpha t/\gamma + \alpha(\alpha+1)(t/\gamma)^2/2 + o(t^3)$ for t near 0. Putting these pieces together, and using $\mathbb{I}$ to denote the indicator function, we obtain the following upper bound on the probability of a change in the number of

good services during the day:

$$
\begin{aligned}
& \mathbb{P}_u[X(\delta) \neq X(0)] \\
\leq\ & \mathbb{P}_u[J^* < \infty] \\
=\ & \sum_{j=1}^{\infty} \mathbb{E}_u[\mathbb{I}[j \leq A]\mathbb{P}_u[J^* = j \mid W_j, W_j(\delta)]] \\
\leq\ & \sum_{j=1}^{\infty} E_u\left[\mathbb{I}[j \leq A]\left(1-\exp[-\bar{\lambda}(u)\xi_j]\right)\right] \\
=\ & \sum_{j=1}^{\infty} \mathbb{P}_u[j \leq A]\mathbb{E}_u\left[1-\exp[-\bar{\lambda}(u)\xi_j]\right] \\
=\ & \sum_{j=1}^{\infty} \mathbb{P}_u[j \leq A]\left(1-\prod_{\ell=1}^{j-1}\mathbb{E}_u[\exp[-\bar{\lambda}(u)\delta S_\ell]]\right) \\
=\ & \sum_{j=1}^{\infty} \mathbb{P}_u[j \leq A]\left(1-(1+\bar{\lambda}\delta/\gamma)^{-(j-1)\alpha}\right) \\
=\ & \sum_{j=1}^{\infty} \mathbb{P}_u[j \leq A]\left[(j-1)\alpha\bar{\lambda}(u)\delta/\gamma+\mathscr{O}(\delta^2 j^2)\right] \\
=\ & \sum_{a=1}^{\infty} \mathbb{P}_u[A=a]\sum_{j=1}^{a}\left[(j-1)\alpha\bar{\lambda}(u)\delta/\gamma+\mathscr{O}(\delta^2 j^2)\right] \\
=\ & \sum_{a=1}^{\infty} \mathbb{P}_u[A=a]\left[\frac{(a-1)a}{2}\frac{\alpha\bar{\lambda}(u)\delta}{\gamma}+\mathscr{O}(\delta^2 a^3)\right] \\
\leq\ & \mathbb{E}_u[A^2]\alpha\bar{\lambda}(u)\delta/\gamma+\mathbb{E}_u[A^3]\mathscr{O}(\delta^2) \\
=\ & \mathscr{O}(\delta b^3).
\end{aligned}
$$

The second equality holds because A and ξ_j are independent. This completes the proof of the lemma. □

Proof of the Proposition 1. Using Holder's inequality with $1/p = 1-1/q$ for an arbitrary $q > 1$ and denoting $\Delta = X(\delta)-X(0)$, we have

$$
\begin{aligned}
\mathbb{E}[\Delta^2 \mid B=b] &\leq \mathbb{E}_u[\Delta^2\mathbb{I}[\Delta^2>0]] \\
&\leq (\mathbb{E}_u[\Delta^{2q}])^{1/q}(\mathbb{E}_u[\mathbb{I}^p[\Delta^2>0]])^{1/p} \\
&\leq (\mathbb{E}_u[A^{2q}])^{1/q}(\mathbb{P}_u[\Delta\neq 0])^{(q-1)/q} \\
&= K_0(q,u)\cdot\mathscr{O}((\delta b^3)^{(q-1)/q}) \\
&= \mathscr{O}((a\cdot b)^2(\delta b^3)^{(q-1)/q})
\end{aligned}
$$

where $(K_0(q,u))^q = \mathbb{E}_u[A^{2q}] = \mathscr{O}((a\cdot b)^{2q})$ is a very loose bound on $E[\Delta^{2q}]$ that does not depend on δ. By selecting $q = 1/\varepsilon$, this gives

$$\mathrm{Var}[\Delta \mid B=b] \leq \mathbb{E}[\Delta^2 \mid B=b] = \mathscr{O}((ab)^2(\delta b^3)^{1-\varepsilon}).$$

On the other hand, since Δ is an integer, we always have $\Delta \leq \Delta^2$ and thus

$$\mathbb{E}[\Delta \mid B=b] \leq \mathbb{E}[\Delta^2 \mid B=b] = \mathscr{O}((ab)^2(\delta b^3)^{1-\varepsilon}).$$

Then,

$$
\begin{aligned}
\mathrm{Var}[\Delta] &= \mathbb{E}[\mathrm{Var}[\Delta \mid B]] + \mathrm{Var}[\mathbb{E}[\Delta \mid B]] \\
&= \mathscr{O}(a^2\mathbb{E}[B^2(\delta B^3)^{1-\varepsilon}]) \\
&\quad + \mathscr{O}(a^4\mathbb{E}[B^4(\delta^2 B^6)^{1-\varepsilon}]).
\end{aligned}
$$

Since B has bounded moments of all orders, this gives $\mathrm{Var}[\Delta] = \mathscr{O}(\delta^{1-\varepsilon})$, which completes the proof. □

The empirical results of the previous subsection indicate that $\mathrm{Var}[\Delta] \in \mathscr{O}(\delta)$, at least for the (a + c) and (b + c) synchronization strategies and with the parameters values of our numerical experiment. If Δ was bounded by a constant K_0, then we would easily have $\mathrm{Var}[\Delta \mid B=b] = K_0^2\mathbb{P}_u[\Delta\neq 0] = K_0^2\mathscr{O}(\delta b^3) = \mathscr{O}(\delta)$. Strictly speaking, Δ is not bounded, but when δ is small, the probability than Δ exceeds a few units is so small that Δ can be considered as bounded from a *practical* viewpoint.

ACKNOWLEDGMENTS

This research has been supported by Grants OGP-0110050 and CRDPJ-251320 from NSERC-Canada, and a grant from Bell Canada via the Bell University Laboratories, to the first author. The second author benefited from an Industrial Scholarship from NSERC-Canada and the Bell University Laboratories.

REFERENCES

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* 50 (7): 896–908.

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

Buist, E., and P. L'Ecuyer. 2005. A Java library for simulating contact centers. In *Proceedings of the 2005 Winter Simulation Conference*, 556–565: IEEE Press.

Ceẓik, M. T., and P. L'Ecuyer. 2006. Staffing multiskill call centers via linear programming and simulation. *Management Science*. To appear.

Cochran, W. G. 1977. *Sampling techniques*. Second ed. New York: John Wiley and Sons.

Gans, N., G. Koole, and A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* 5:79–141.

Glasserman, P. 1991. *Gradient estimation via perturbation analysis*. Norwell, MA: Kluwer Academic.

Glasserman, P., and D. D. Yao. 1992. Some guidelines and guarantees for common random numbers. *Management Science* 38 (6): 884–908.

Glynn, P. W., and R. Szechtman. 2002. Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, ed. K.-T. Fang, F. J. Hickernell, and H. Niederreiter, 27–49. Berlin: Springer-Verlag.

Lavenberg, S. S., and P. D. Welch. 1981. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science* 27:322–335.

L'Ecuyer, P. 1990. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science* 36 (11): 1364–1383.

L'Ecuyer, P. 2006. *Stochastic simulation*. Notes for a graduate simulation course.

L'Ecuyer, P., and E. Buist. 2005. Simulation in Java with SSJ. In *Proceedings of the 2005 Winter Simulation Conference*, 611–620: IEEE Press.

L'Ecuyer, P., and G. Perron. 1994. On the convergence rates of IPA and FDC derivative estimators. *Operations Research* 42 (4): 643–656.

Whitt, W. 1999. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* 24:205–212.

AUTHOR BIOGRAPHIES

PIERRE L'ECUYER is Professor in the Département d'Informatique et de Recherche Opérationnelle, at the Université de Montréal, Canada. He holds the Canada Research Chair in Stochastic Simulation and Optimization. His main research interests are random number generation, quasi-Monte Carlo methods, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems, and discrete-event simulation in general. He is currently Associate/Area Editor for *ACM TOMACS*, *ACM TOMS*, and *Statistical Computing*. He obtained the prestigious *E. W. R. Steacie* fellowship in 1995-97 and a *Killam* fellowship in 2001-03. His recent research articles are available on-line from his web page: <http://www.iro.umontreal.ca/~lecuyer>.

ERIC BUIST is a PhD Student at the Université de Montréal. His main interests are software engineering, object-oriented programming, and simulation. He is currently working on the development of flexible and efficient tools for the simulation of contact centers. His e-mail address is <buisteri@IRO.UMontreal.CA>.