2005

Performance Measures for Service Systems with a Random Arrival Rate

Vijay Mehrotra
University of San Francisco, vmehrotra@usfca.edu

Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds

PERFORMANCE MEASURES FOR SERVICE SYSTEMS WITH A RANDOM ARRIVAL RATE

Samuel G. Steckley and Shane G. Henderson

School of OR & IE
Cornell University
Ithaca, NY 14853, U.S.A.

Vijay Mehrotra

Decision Sciences Department
San Francisco State University
San Francisco, CA 94132, U.S.A.

ABSTRACT

It is commonly assumed that the arrival process of customers to a service system is a nonhomogeneous Poisson process. Call center data often refute this assumption, and several authors have postulated a doubly-stochastic Poisson process for arrivals instead. We develop approximations for both the long-run fraction of calls answered quickly, and the distribution of the fraction of calls answered quickly within a short period. We also perform a computational study to evaluate the approximations and improve our understanding of such systems.

1 INTRODUCTION

Workforce management in telephone call centers is largely dependent on strong workforce planning and management (Mehrotra 1997, Cleveland and Mayben 1997), as typically 60-80% of a call center's budget is spent on the labor costs associated with the agents who handle customer phone calls. Conversely, it is well documented that in many call center environments there are significant hidden costs and risks associated with delivering service quickly (Ittig 1994, Ittig 2002, Pullman and Moore 1999).

Call center workforce planning and management involves three levels of analysis and decision-making:

- "Long Run" Planning (typically 6-12 months in advance),
- "Short Term" Scheduling (typically 1-2 weeks in advance), and
- "Real Time" Schedule Adjustments

All of these workforce management processes depend explicitly on the ability to accurately translate demand for service (measured in terms of call volumes and call handling times) into a demand for agents (which depends on waiting time distribution objectives defined by management as well as forecasted workload).

The conventional approach to this translation is to model call queues as "Stationary, Independent Period-by-Period," (SIPP) as described in Green et al. (2001). The SIPP approach divides the planning horizon into a series of periods (time intervals), e.g., Monday 8 - 8:30am. Within each period a stationary queueing model is analyzed to provide estimates of performance in that period.

The arrival processes within periods are usually modeled as independent Poisson Processes, with an arrival rate that is assumed to be deterministic and fixed throughout the period. Agent requirements for each period are then determined from steady state equations that are based on the forecasted arrival and service rates, and target service objective for that period.

There are a number of potentially significant problems associated with the SIPP approach. We believe that the most significant problem is the assumption of a deterministic arrival rate for each period. Recent empirical studies (notably Brown et al. 2005 and Avramidis et al. 2004) have suggested that there is often significant variability in call center arrival rates. We show that if this variability is not accounted for in the determination of the number of servers, then understaffing and poor service quality (that is, long customer waiting times and high abandonment rates) can result.

In this paper we examine the impact of randomly varying arrival rates on call center system performance. In particular, we compute performance approximations for the case where, in each instance of a particular period, the arrival rate is first sampled from a distribution, and then arrivals in the period occur according to a homogeneous Poisson process with that arrival rate.

Our performance approximations are related to the fraction of calls answered within a given time limit. In contrast, Harrison and Zeevi (2005) and Whitt (2004) adopt an "economic" model where costs are assumed

for abandoned calls, waiting times of customers and agents. They then minimize cost over the choice of staffing levels.

It is important to distinguish the randomly-varying arrival rate behavior discussed in this paper from the randomness arising from forecast uncertainty. In such a case, the arrival rate for a period is assumed to be a deterministic quantity, but it is not known with certainty. This case often arises in one-time planning, when, for example, a new product is introduced. The appropriate long-run performance measure may differ in this case, as discussed in Steckley, Henderson, and Mehrotra (2005). (This case is called the "uncertain arrival rate" case in that paper.) We do not discuss that case further here.

In Section 2, we develop our approximations for long-run performance, and the distribution of performance in a single period. In Section 3, we describe a set of experiments designed to evaluate the quality of our performance approximations. We then compare these performance approximations with simulation results and explain the observed trends. Conclusions and suggestions for future research are presented in Section 4.

To better understand the model of call arrivals we treat in this paper we describe a particular example originally proposed in Whitt (1999). In this model, the arrival process on a given day is Poisson with arrival rate function $B(\lambda(s) : s \geq 0)$, where $(\lambda(s) : s \geq 0)$ is a "profile" describing the relative intensities of arrivals, and B is a random "busyness" parameter indicating how busy the day is. To simplify the analysis we assume that $\lambda(\,)$ is constant within each period. We use this model for the experiments in Section 3 but the analytic results in Section 2 do not rely on this particular choice of model.

2 COMPUTING PERFORMANCE WITH RANDOMLY VARYING ARRIVAL RATES

For a given period the key long-run performance measure is the long-run fraction of customers that receive satisfactory service. A customer receives satisfactory service if her delay in queue is at most τ seconds. Common choices for τ are 20 seconds (a moderate delay) and 0 seconds (no delay). For much of what follows we focus on a single period (e.g., 10am - 10:15am) in the day, arbitrarily representing this time period as time 0 through time t. Let Λ_i denote the real-valued random arrival rate within this period on day i. (A period may only occur once each week, such as the period Monday from 8 - 8:15am. In this case, the term "day i" should be interpreted as the ith realization of the period.)

Let S_i denote the number of satisfactory calls (calls that are answered within the time limit τ) out of a total of N_i calls that are received in the period on day i. Notice that here we consider any call that abandons to be unsatisfactory. Some planners prefer to ignore calls that abandon within very short time frames. There is a difference, but it is not important for our discussion.

Over n days, the fraction of satisfactory calls is

$$\frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} N_i}$$

Assume that days are i.i.d., the staffing level is fixed throughout, and $EN_1 < \infty$. (Assuming days are i.i.d. ignores the inter-day correlations seen in Brown et al. (2005) and Steckley, Henderson, and Mehrotra (2005). More general dependence structures can be captured in essentially the same framework.) The last assumption holds if $E\Lambda_1 < \infty$. Dividing both the numerator and denominator by n and taking the limit as $n \to \infty$, the strong law then implies that the long-run fraction of satisfactory calls is

$$\frac{ES_1}{EN_1} \tag{1}$$

This ratio gives performance as a function of staffing level. But how do we compute it?

First note that

$$\begin{aligned} EN_1 &= EE[N_1|\Lambda_1] \\ &= E[\Lambda_1 t] \\ &= tE\Lambda_1, \end{aligned} \tag{2}$$

so that EN_1 is easily computed. Computing ES_1 is more difficult. We again condition on Λ_1 to obtain $ES_1 = Es(\Lambda)$, where $s(\lambda)$ is the conditional expected number of satisfactory calls in the period, conditional on $\Lambda_1 = \lambda$. Our initial goal is an expression for $s(\lambda)$.

Fix the arrival rate to be deterministic and equal to λ (for now). Let $X(\,;\lambda) = (X(s;\lambda) : s \geq 0)$ be a Markov process used to model the call center when there is a fixed arrival rate λ. In specialized cases one can take X to be the process giving the number of customers in the system, but it may be more complicated. Suppose that a customer arriving at time s will receive satisfactory service if and only if $X(s;\lambda) \in B$ for some distinguished set of states B.

Example 1 *A common model of a call center is an $M/M/c+M$ queue, i.e., the Erlang-A model. There are c servers, service times are exponentially distributed, and the arrival process is Poisson. Customers are willing to wait an exponentially-distributed amount of time (the "patience time") in the queue, and abandon if they do not reach a server by that time. Here we take $X(s;\lambda)$ to be the number of customers in the system at time s. Then X is a continuous-time Markov chain (CTMC).*

Suppose that a service is considered satisfactory if and only if the customer immediately reaches a server. Then we can take $B = \{0, 1, 2, \ldots, c-1\}$, i.e., a service is satisfactory if and only if the number of customers in the system is $c-1$ or less when the customer arrives.

Example 2 *Consider the same model as in the previous example, but now define a service to be satisfactory if and only if the customer reaches a server in at most $\tau > 0$ seconds so long as she doesn't abandon. The state space of the CTMC defined in the previous example is no longer rich enough to determine, upon a customer arrival, whether that customer will receive satisfactory service or not. We turn to a different Markov process in such a case. Without loss of generality, suppose that as soon as a customer arrives, the patience and service times for that customer are sampled and therefore known. Since customers are served in FIFO order we can determine, for every customer that has arrived by time s, whether that customer will abandon or not, and if not which agent the customer will be served by. Let $V_i(s;\lambda)$ denote the "work in process" for agent i at time s, $i = 1, \ldots, c$. The quantity $V_i(s;\lambda)$ gives the time required for agent i to complete the service of all customers in the system at time s that are, or will be, served by agent i. Let $X(s;\lambda)$ be the vector $(V_i(s;\lambda) : 1 \le i \le c)$. The process $X(\,;\lambda) = (X(s;\lambda) : s \ge 0)$ is a Markov process, albeit a rather complicated one, and we can take $B = \{v : \min_{i=1}^{c} v_i \le \tau\}$, so that a service is satisfactory if and only if at least one server will be available to answer a call within τ seconds of a customer's arrival.*

Let $P_\varphi(\,)$ denote the probability measure when the Markov process has initial distribution φ. Let ν and π be, respectively, the distribution of the Markov process at time 0 and the stationary distribution (assumed to exist and be unique). Proposition 1 serves as a foundation for the use of steady-state approximations for performance measures in both the deterministic and random arrival rate contexts. The proof of this result is based on an application of "Poisson arrivals see time averages" results; see Steckley, Henderson, and Mehrotra (2005).

Proposition 1 *Under the conditions above,*

$$s(\lambda) = \lambda \int_0^t P_\nu(X(s;\lambda) \in B)\, ds$$

If $\nu = \pi$, so that the Markov process is in steady-state at time 0, then

$$s(\lambda) = \lambda t f(\lambda),$$

where $f(\lambda) = P_\pi(X(0;\lambda) \in B)$ is the steady-state probability that the system is in state B. We can interpret $f(\lambda)$ as the long-run fraction of customers that receive satisfactory service.

2.1 Steady-state approximations

Suppose that we adopt the steady-state approximation $s(\lambda) \approx \lambda t f(\lambda)$. Here λt is the expected number of customer arrivals in the period and $f(\lambda)$ is the long-run fraction of customers that receive satisfactory service. From (1) and (2), we see that

$$\frac{\mathrm{E}S_1}{\mathrm{E}N_1} = \frac{\mathrm{E}s(\Lambda_1)}{t\mathrm{E}\Lambda_1} \approx \frac{\mathrm{E}[\Lambda_1 f(\Lambda_1)]}{\mathrm{E}\Lambda_1} \tag{3}$$

The fact that one should weight $f(\Lambda)$ by the arrival rate in (3) is well known. It is implicit (and at times explicit) in the work of Harrison and Zeevi (2005) and Whitt (2004) for example. Chen and Henderson (2001) did *not* perform this weighting in their analysis, so their results do not directly apply to the RVAR case, in contrast to what is claimed there.

What are the consequences of ignoring a randomly-varying arrival rate when predicting performance in a call center? In that case we would first estimate a deterministic arrival rate. The most commonly used estimates converge to $\mathrm{E}\Lambda_1$ as the data size increases. We then estimate performance as $f(\mathrm{E}\Lambda_1)$.

Together with (3), Proposition 2 below establishes that if f is decreasing and concave over the range of Λ_1, then we will overestimate performance if a random arrival rate is ignored. The function f is, in great generality, decreasing in λ. For many models it is also concave, at least in the region of interest; see Chen and Henderson (2001).

Proposition 2 *Suppose that f is decreasing and concave on the range of Λ_1. Then*

$$\frac{\mathrm{E}[\Lambda_1 f(\Lambda_1)]}{\mathrm{E}\Lambda_1} \le f(\mathrm{E}\Lambda_1)$$

Proof: We have that

$$\begin{aligned} \mathrm{E}[\Lambda_1 f(\Lambda_1)] &\le (\mathrm{E}\Lambda_1)(\mathrm{E}f(\Lambda_1)) && (4)\\ &\le (\mathrm{E}\Lambda_1) f(\mathrm{E}\Lambda_1) && (5) \end{aligned}$$

establishing the result. The inequality (4) follows since f is decreasing (see, e.g., Whitt 1976), and (5) uses Jensen's inequality. □

For certain models and distributions of Λ_1, we may be able to compute (3) exactly. In general though, this will not be possible. In such a case we can use some numerical integration technique. The problem is quite straightforward since f is typically easily computed and the integral $\mathrm{E}[\Lambda_1 f(\Lambda_1)]$ is one-dimensional.

We now turn from long-run performance to short-run performance. We want to determine the distribution of S_1/N_1, the fraction of satisfactory calls in a single

period $[0,t]$ of a single day (We define $0/0 = 1$.) Our approach is to condition on Λ, the arrival rate for the period.

Suppose that conditional on Λ, the period is long enough that the fraction of calls answered on time is close to its steady-state mean $f(\Lambda)$. This transformation of the random variable Λ is our first approximation. It ignores the "process variability" that arises even for a fixed arrival rate.

We can refine this approximation to take into account process variability. The key to the refinement is a central limit theorem (CLT) for S_1/N_1 assuming a fixed λ. The CLT should hold in great generality, as argued in Steckley, Henderson, and Mehrotra (2005). Here we establish the CLT under strong conditions, and provide a computable expression for the variance in the process.

Let the arrival rate λ be fixed. Suppose that our goal is to answer calls immediately. Suppose further that the number-in-system process $X = (X(s) : s \geq 0)$ can be modeled as an irreducible continuous-time Markov chain on the finite state space $\{0, 1, \ldots, d\}$, where $d > c$. (It is not essential that the state space be finite, but it allows us to avoid verifying regularity conditions.) Let $M(s)$ be the number of transitions by time s, and let $Y = (Y_n : n \geq 0)$ be the embedded discrete-time Markov chain. Then we can write

$$\frac{S_1}{N_1} \approx \frac{U_{M(t)}}{V_{M(t)}}, \tag{6}$$

where

$$\begin{aligned} U_n &= \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1} + 1, Y_{i-1} \leq c-1) \text{ and} \\ V_n &= \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1} + 1) \end{aligned}$$

Here U_n gives the fraction of the first n transitions that correspond to an arriving customer finding a server available. Similarly, V_n gives the fraction of the first n transitions that correspond to an arrival joining the system. Notice that V_n does not count blocked customers. This is why the relation in (6) is not an equality. When d is large enough that few customers are turned away, the approximation should be very good.

Theorem 1 *Under the assumptions given above,*

$$\sqrt{\lambda s}\left(\frac{U_{M(s)}}{V_{M(s)}} - \frac{u}{v}\right) \Rightarrow N(0, \sigma^2(\lambda))$$

as $s \to \infty$, where u, v and $\sigma^2(\lambda)$ are specified in the proof below.

Proof: The proof has 3 steps. The key step is to establish the joint CLT

$$\sqrt{n}\left(\begin{pmatrix} U_n \\ V_n \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow N(0, \Sigma) \tag{7}$$

as $n \to \infty$, where $N(0,\Sigma)$ denotes a Gaussian random vector with mean 0 and covariance matrix Σ, and u, v and Σ are specified below. The final 2 steps consist of applying a random time change and then the delta method.

To establish (7) we apply a Markov chain CLT (see, e.g., Meyn and Tweedie (1993), Theorem 17.4.4). That result applies only to univariate processes, but the result easily extends to multivariate processes through an application of the Cramér-Wold device (see, e.g., Billingsley (1968), Theorem 7.7). Consider the (irreducible, finite-state-space) Markov chain $\tilde{Y} = (\tilde{Y}_i : i \geq 0)$, where $\tilde{Y}_i = (Y_i, Y_{i+1})$. We can write

$$\begin{aligned} U_n - u &= \frac{1}{n}\sum_{i=0}^{n-1} \tilde{h}_1(\tilde{Y}_i) \text{ and} \\ V_n - v &= \frac{1}{n}\sum_{i=0}^{n-1} \tilde{h}_2(\tilde{Y}_i), \end{aligned}$$

where

$$\begin{aligned} \tilde{h}_1(x,y) &= I(y = x+1, x \leq c-1) - u \text{ and} \\ \tilde{h}_2(x,y) &= I(y = x+1) - v \end{aligned}$$

Let $\tilde{\pi}$ be the stationary distribution of $\tilde{Y}$. We choose u and v to be steady-state means, so that $\tilde{\pi}\tilde{h}_i = \sum_{(x,y)} \tilde{\pi}(x,y)\tilde{h}_i(x,y) = 0$ for $i = 1, 2$. Let $\tilde{P}$ be the transition matrix of $\tilde{Y}$, and let $\tilde{g}_1$ and $\tilde{g}_2$ solve Poisson's equation

$$\tilde{P}\tilde{g}_i(x,y) = \tilde{g}_i(x,y) - \tilde{h}_i(x,y),$$

for $i = 1, 2$ and all (x, y). We then obtain (7), where

$$\begin{aligned} \Sigma_{ij} &= \mathrm{E}_{\tilde{\pi}}[(\tilde{g}_i(\tilde{Y}_1) - \tilde{P}\tilde{g}_i(\tilde{Y}_0))(\tilde{g}_j(\tilde{Y}_1) - \tilde{P}\tilde{g}_j(\tilde{Y}_0))] \\ &= \mathrm{E}_{\tilde{\pi}}[\tilde{g}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0) + \tilde{h}_i(\tilde{Y}_0)\tilde{g}_j(\tilde{Y}_0) \\ &\quad -\tilde{h}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0)] \end{aligned}$$

The second equality follows as in Meyn and Tweedie (1993), Equation 17.47.

In fact, we obtain a stronger result, namely a functional CLT. This latter observation, together with the random-time-change result (Billingsley 1968, Theorem

17 1), allows us to conclude that

$$\sqrt{M(s)}\left(\begin{pmatrix} U_{M(s)} \\ V_{M(s)} \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow N(0,\Sigma)$$

as $s \to \infty$ Now, $M(s)/s \to \gamma$ as $s \to \infty$ a s , where γ is the long-run rate of transitions in the continuous-time Markov chain X The converging-together lemma (Billingsley 1968, Problem 1, p 28) then implies that

$$\sqrt{\gamma s}\left(\begin{pmatrix} U_{M(s)} \\ V_{M(s)} \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow N(0,\Sigma)$$

as $s \to \infty$

The final step applies the delta method (e g , Billingsley (1968), Problem 2, p 34, using the function $\phi(x,y) = x/y$, to conclude that

$$\sqrt{\gamma s}\left(\frac{U_{M(s)}}{V_{M(s)}} - \frac{u}{v}\right) \Rightarrow N(0,\eta^2),$$

where

$$\begin{aligned} \eta^2 &= \nabla\phi(u,v)^T \Sigma \nabla\phi(u,v) \\ &= \frac{\Sigma_{11} - 2(u/v)\Sigma_{12} + (u/v)^2\Sigma_{22}}{v^2} \end{aligned}$$

Setting $\sigma^2(\lambda) = \lambda\eta^2/\gamma$ yields the result □

Equation (6) and Theorem 1 establish that conditional on Λ, the fraction S_1/N_1 is approximately normally distributed with mean $f(\Lambda)$ and variance $\sigma^2(\Lambda)/\Lambda t$ So we can approximate the distribution of S_1/N_1 by the normal mixture $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$

Remark 1 The variance of this normal mixture is

$$\text{Var } f(\Lambda) + \text{E}\frac{\sigma^2(\Lambda)}{\Lambda t},$$

which can be viewed as a decomposition of the variance into contributions from arrival rate uncertainty and process uncertainty respectively

To compute the distribution of this normal mixture we need to be able to compute the constant $\sigma^2(\lambda)$, which in turn depends on γ and η^2 (which also depend on λ) The following formulae are useful in this regard They exploit the strong relationships between the 2-step Markov chain $\tilde{Y}$ and the single-step Markov chain Y, and between the continuous-time Markov chain X and its embedded chain Y Let $\beta(i)$ denote the rate at which the CTMC X leaves state i, and let π_X and π_Y denote the steady-state distributions associated with X and Y respectively Since

$$\pi_X(y) = \frac{\pi_Y(y)/\beta(y)}{\sum_z \pi_Y(z)/\beta(z)},$$

it follows that

$$\gamma = \sum_{y=0}^{d} \pi_X(y)\beta(y) = \left(\sum_{z=0}^{d} \pi_Y(z)/\beta(z)\right)^{-1}$$

Note that π_X or π_Y are easily computed, and therefore so is γ

We also need to compute u and v These are given by

$$\begin{aligned} u &= \sum_{i=0}^{c-1} \pi_Y(i) P_Y(i,i+1) \text{ and} \\ v &= \sum_{i=0}^{d-1} \pi_Y(i) P_Y(i,i+1), \end{aligned}$$

where P_Y is the transition matrix of Y

Finally, recall that for $1 \le i,j \le 2$

$$\begin{aligned} \Sigma_{ij} &= \text{E}_{\tilde{\pi}}[\tilde{g}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0) + \tilde{h}_i(\tilde{Y}_0)\tilde{g}_j(\tilde{Y}_0) \\ &\quad -\tilde{h}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0)] \\ &= \sum_{x,y} \pi_Y(x) P_Y(x,y)[\tilde{g}_i(x,y)\tilde{h}_j(x,y) \\ &\quad +\tilde{h}_i(x,y)\tilde{g}_j(x,y) - \tilde{h}_i(x,y)\tilde{h}_j(x,y)] \end{aligned}$$

It remains to specify how to compute $\tilde{g}_i(x,y)$ Define

$$h_i(x) = \text{E}_x\tilde{h}_i(x,Y_1) = \sum_{y=0}^{d} \tilde{h}_i(x,y) P_Y(x,y)$$

to be the "smoothed" version of $\tilde{h}_i$, for $i = 1,2$ and $x = 0, \quad , d$ There are multiple solutions to the equations defining $\tilde{g}_i$, all of which differ by an additive constant In what follows we use one such solution for $\tilde{g}_i$, which is

$$\begin{aligned} \tilde{g}_i(x,y) &= \sum_{k=0}^{\infty} \text{E}_{(x,y)}\tilde{h}_i(Y_k, Y_{k+1}) \\ &= \tilde{h}_i(x,y) + \sum_{k=1}^{\infty} \text{E}_{(x,y)}\tilde{h}_i(Y_k, Y_{k+1}) \\ &= \tilde{h}_i(x,y) + \sum_{k=1}^{\infty} \text{E}_{(x,y)} h_i(Y_k) \\ &= \tilde{h}_i(x,y) + g_i(y), \end{aligned}$$

where

$$g_i(y) = \sum_{k=0}^{\infty} \mathrm{E}_y h_i(Y_k)$$

solves $(P_Y - I)g_i(y) = -h_i(y)$ for all y, and has the property that $\pi_Y g_i = 0$. It is therefore possible to compute g_i from these latter relations, and then substitute back to obtain $\tilde{g}_i$.

2.2 Simulation-based estimates

The approximations for long-run and short-run performance described above may be inappropriate, either because the steady-state approximations for time-dependent quantities may be inaccurate for a non-negligible set of arrival rates, or because the true system is not well modelled by simple models for which steady-state results are readily computed. It is natural to then turn to simulation to compute performance measures.

In terms of long-run performance, we have already noted that the problem reduces to computing $\mathrm{E}S_1$, the expected number of satisfactory calls in a particular period. This is straightforward using simulation. One can simply generate the arrival rate process, Λ say, and then conditional on the realized value, simulate the call center for the day, giving a realization of S_1. Repeating this process in i.i.d. fashion gives $S_1, \ldots, S_n$ say, which can be averaged to give an estimate of $\mathrm{E}S_1$.

For short-run performance we wish to compute the distribution of S_1/N_1. This random variable does not have a (Lebesgue) density since it is supported on the rationals. Its probability mass function is also uninformative. Therefore, we would probably estimate a moderately coarse histogram (say, with bins of width $\Delta x = 0.01$). The height of the bin $[x, x + \Delta x]$ is proportional to $F(x + \Delta x) - F(x)$, where F is the distribution function of S_1/N_1. Hence, estimating this histogram is equivalent to estimating the distribution function at the fixed set of points $\Delta x, 2\Delta x, \ldots, 1$. This estimation is straightforward based on i.i.d. observations (S_1, N_1), and one can apply standard results (e.g., Ross 1996, pp. 360–363) to compute tolerance bounds for F.

3 EXPERIMENTAL INSIGHTS

We conducted experiments to examine performance given uncertainty in the arrival rate. Specifically, we wanted to determine which factors impact the performance measures discussed in §2, assess the quality of the approximations as compared to the simulation-based estimates of performance, and learn more about the behavior of systems with a random arrival rate. The factors we chose to examine included (a) the level of variability in the (Poisson) arrival rate; (b) the duration of the (exponential) service times; and (c) the (exponential) rate at which customers abandon the system.

Note that we continue to focus our analysis on a single period. The design of the experiment is discussed in §3.1 and the results are presented in §3.2.

3.1 Experimental Design

For our experiments, we model the call center as an $M/M/c + M$ queue (i.e., the Erlang-A model) with a random arrival rate Λ. We adopt the Whitt (1999) model discussed earlier in which the arrival rate in the ith instance of the period is given by $B_i\lambda$, where the B_is are i.i.d. We model B_i as uniform with mean 1 so that Λ is uniform with mean λ. We chose the uniform distribution because it is simple and it effectively illustrates the essential ideas. One could easily substitute a more realistic distribution. The choice of the endpoints of the uniform distribution are discussed below.

For these experiments, we have set the length of the period at one hour. A call is defined to have received satisfactory service if it is answered immediately, i.e., $\tau = 0$.

Using both the analytic approximations discussed above and the corresponding simulation models, we estimate the performance measures discussed in §2.1 and §2.2 for a number of scenarios. The simulations were modelled and run using software developed by Eric Buist and Pierre L'Ecuyer (Buist and L'Ecuyer 2005), which was chosen for its ease of modeling call center operations and capturing the desired performance statistics, as well as its very fast simulation run times.

The scenarios are summarized in Table 1. We vary the expected number of calls per hour (λ). We also vary the variability in the arrival rate in terms of a quantity we call the variance factor. The variance factor is defined as the ratio of the variance of the number of calls per hour under the random arrival rate Λ and the variance of the number of calls per hour given a deterministic arrival rate λ. The level of the variance factor then determines the endpoints of the uniform distribution for Λ and thus determines the variability of Λ. Finally, we allow the mean service time and mean abandonment time to vary.

The range of variance factors (as well as arrival rates and average handle times) included in these experiments is based on the actual historical data from four diverse call centers that we have studied; additional details and examples from this dataset are presented in Steckley, Henderson, and Mehrotra (2005).

In Table 1 a variance factor of one corresponds to the case in which the arrival rate is deterministic and equal to λ. An abandonment rate of 0 corresponds to

the case in which there is no abandonment, in which case the call center is modeled as an $M/M/c$ queue

Table 1: Experimental Design

Factor	Levels
Mean number of calls per hour (λ)	250 1000 4000
Variance factor	1 3 6
Service rate per hour (μ)	12 6
Abandonment rate per hour(θ)	0 6 12

For each scenario, we selected the number of servers c to be the minimum value so that the long-run fraction of calls that are served immediately for a system with a deterministic arrival rate λ is at least 90%.

For the simulations, we used an extensive warm-up period. The parameter settings (arrival rate, service time distribution, abandonment time distribution) for the warm-up period were identical to those used in the simulation of the actual period for which data was captured. Therefore, our data reflects steady-state performance.

3.2 Results

Both the simulation-based estimates and steady-state approximations for long-run performance (long-run fraction of satisfactory calls) are reported in Table 2. The simulation results are accurate to approximately 2 decimal places, and so are reported only to that accuracy. Due to space considerations we present only selected scenarios. This selection illustrates the essential characteristics and trends seen in the results as a whole.

The approximations and simulation-based estimates are very similar. We expect such agreement since the simulated period should exhibit steady-state behavior after the extensive warm-up we used.

When the variance factor is one so that there is no variability in the arrival rate, the long-run fraction of satisfactory calls is very close to 0.9. This is because the number of servers c is specifically chosen so that the long-run fraction of satisfactory calls will be at least 0.9 in this case. When the variance factor is strictly greater than one, so that there is variability in the arrival rate, the long-run fraction of satisfactory calls is less than 0.9 as suggested by Proposition 2. We also see that the more variable the arrival rate, the worse the performance. We see that the degradation can be significant. It is on the order of 5% - 10% for some of the cases.

Table 2: Simulation-Based Estimates and Approximations (in Parentheses) of Long-Run Performance

			Variance factor		
λ	μ	θ	1	3	6
250	12	0	0.91 (0.91)	0.87 (0.87)	0.82 (0.81)
1000	12	0	0.90 (0.90)	0.87 (0.87)	0.82 (0.82)
4000	12	0	0.91 (0.90)	0.88 (0.87)	0.83 (0.82)
1000	6	0	0.91 (0.91)	0.85 (0.84)	0.76 (0.76)
1000	12	6	0.89 (0.90)	0.87 (0.87)	0.84 (0.84)
1000	12	12	0.90 (0.91)	0.88 (0.89)	0.86 (0.86)

The results also indicate that abandonment reduces the negative impacts of variability in the arrival rate. To understand this, note that in a no-abandonment model, customers with long waiting times remain in the system, creating a "chain reaction" of waiting for future customers. In contrast, with abandonment, these customers leave the system quickly, thereby avoiding the chain reaction encountered in a no-abandonment model. This reasoning suggests that the same trend would be observed if we had instead defined a call to have received satisfactory service if the call does not abandon and is answered within $\tau > 0$ seconds. Although we believe this trend holds in general, in some cases in which τ is very large and the rate of abandonment θ is also very large, the abandoning calls may actually drive down the long-run fraction of satisfactory calls.

For short-run performance, we turn to the distribution of S_1/N_1, the fraction of satisfactory calls in a single instance of the period. We have two possible approximations for this distribution. The first is given by the distribution of $f(\Lambda)$. The second is given by the distribution of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$. Figure 1 plots the simulation-based estimate of the distribution (histogram) along with the the density of the two approximations for a particular case. The final bar of the histogram corresponds to the observed S_1/N_1 ratios that were exactly one. The density of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$ has been truncated at one and the probability of the truncated region has been plotted as a "histogram" bar just to the right of one. The density of $f(\Lambda)$ is obtained by smoothing a histogram of $f(\Lambda)$, which explains its slightly irregular appearance.

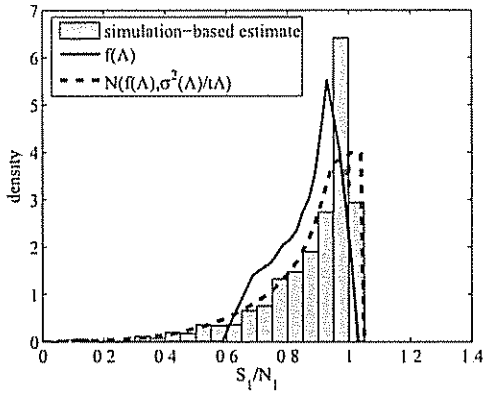

Figure 1: Plots of the Distribution Estimates when $\lambda = 1000$, $\mu = 12$, $c = 97$, $\theta = 0$, and the Variance Factor = 3

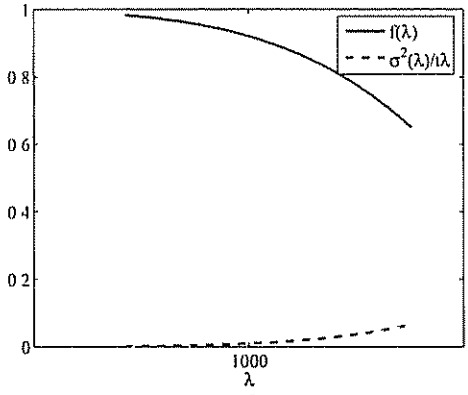

Figure 2: Plot of $f(\,)$ and $\sigma^2(\,)/t(\,)$ for the Scenario of Figure 1

The simulation-based histogram shows that the distribution of S_1/N_1 has a spike around one and a skewed left tail for the given staffing level. We saw the same general shape for all the scenarios in which there is variability in the arrival rate. The shape indicates that it is quite likely that performance for a single instance of the period will be excellent with the fraction of satisfactory calls greater than 0.9. But with a significant probability, the fraction of satisfactory calls will be less than 0.9 and can be as bad as 0.5.

The approximations in Figure 1 track the simulation-based results fairly well. The normal mixture approximation is a much better estimate in the left tail.

To better understand the general shape of the distribution when there is variability in the arrival rate, consider Figure 2 which plots the mean $f(\,)$ and variance $\sigma^2(\,)/(\,)t$ of the normal mixture over the support of the arrival rate distribution for the case plotted in Figure 1. When the arrival rate is small, the mean is very close to one and the variance is very small. This corresponds to the situation in which the call center is comfortably overstaffed and nearly all calls receive satisfactory service. For such λ, $N(f(\lambda), \sigma^2(\lambda)/\lambda t)$ has a very concentrated density in the neighborhood of one. The larger arrival rates result in lower means and higher variances. This corresponds to a situation in which the call center is understaffed and performance becomes more variable. In such cases, $N(f(\lambda), \sigma^2(\lambda)/\lambda t)$ takes on small values and is more dispersed.

In Figure 3, we present a plot of the various estimates for the case in which all parameters are the same, except the variance factor which has increased to 6. There is now an even greater skew in the left tail, which means

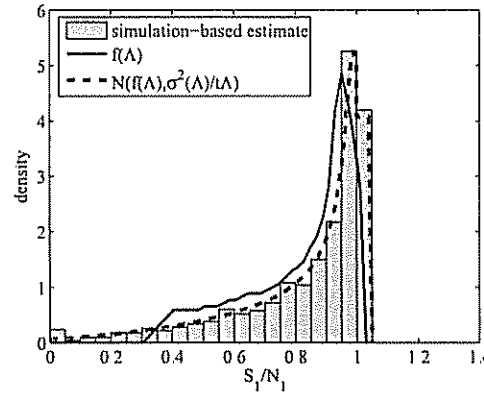

Figure 3: Plots of the Distribution Estimates when $\lambda = 1000$, $\mu = 12$, $c = 97$, $\theta = 0$, and the Variance Factor = 6

that there is higher probability of disastrous performance for a single instance of a period. In fact, as variability in the arrival rate becomes extremely large (variance factor ≥ 50), the distribution of S_1/N_1 becomes bimodal with one mode at 1 and the other at 0. Intuitively, the arrival rate distribution is so spread out that it rarely takes on values that our staffing level is designed to handle, instead taking values that are either very large, or very small relative to the staffing level. Therefore, performance is either very poor, or very good, with little chance of moderate performance.

Further examination of Figures 1 and 3 suggests that the approximations improve as variability in the arrival rate increases. Indeed, we saw this trend in the other scenarios in our experimental design. To understand this trend, first note that the normal approximation for S_1/N_1 is provably good when the periods are long,

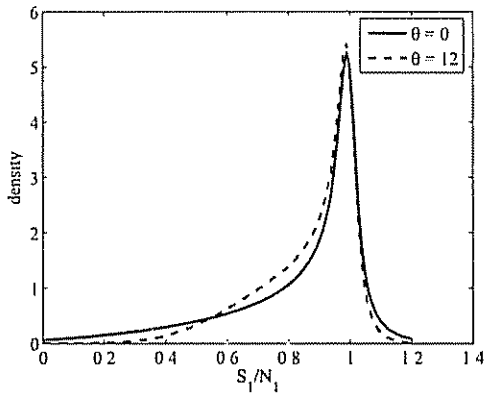

Figure 4: Plot of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$ when $\theta = 0$ and $\theta = 12$, with $\lambda = 1000$, $\mu = 12$, $c = 97$, and the Variance Factor = 6

but deteriorates as the periods become shorter. For shorter periods, N_1 can be small with high probability. As a consequence, the actual distribution of S_1/N_1 will exhibit a right skew. Note that the right skew will be less for small λ since S_1/N_1 then clusters around one. But for any deterministic λ, there will be a discrepancy in the symmetric normal approximation and the right-skewed actual distribution. When the arrival rate Λ is random we smooth the normal approximation over the possible values of Λ to get our approximation $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$. The approximation is essentially a kernel density estimate with local bandwidth $\sigma^2(\,)/t(\,)$. Figure 2 shows that for large λ, where the discrepancy between the normal approximation and actual distribution is significant, $\sigma^2(\lambda)/t(\lambda)$ is relatively large and we smooth more heavily. For smaller λ when the discrepancy is less significant, we do less smoothing. As a result, the approximation gets visually tighter.

To examine the effect of abandonment on short-run performance, we plot the density of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$ for a particular scenario with, and without, abandonment in Figure 4. The densities are very similar around one but the density corresponding to abandonment is less skewed to the left. Similar characteristics are seen in the simulation-based histogram and the distribution of $f(\Lambda)$. The intuition here is the same as for the effect of abandonment on long-run performance.

4 CONCLUSIONS

We have developed approximations for both long-run performance (a single number) and short-run performance (a distribution), where performance is measured in terms of the fraction of calls answered within a reasonable time frame. The long-run approximations perform very well. The short-run approximations are good, and improve as the variability in the arrival rate increases. For many parameter regimes it is important to take into account the process variability exhibited through the function $\sigma^2(\,)$. Not doing so leads to underestimation of the tail behavior.

The short-run performance measures provide valuable information to managers, partly because they clarify the variability in performance that one might expect in a single period: We expect good periods and bad periods, and our results quantify *how often* good and bad periods will arise. They are also valuable because financial contracts are often based on short-run performance figures, and therefore the *distribution* of short-run performance is extremely important.

Several avenues for future research suggest themselves.

- Service times are often better modeled as lognormal random variables than exponential random variables. For such cases, can one obtain exact values or approximations for the mean and variance functions f and σ^2?
- Can one obtain the functions f and σ^2 when performance is measured instead as the fraction of calls that are answered within $\tau > 0$ seconds? (We only treated the $\tau = 0$ case.)
- How does employee absenteeism fit into this framework? Presumably, with a random number of servers, in addition to a random arrival rate, the strong-law approximation would be even better. This seems to be the view of Harrison and Zeevi (2005) and Whitt (2004) who use fluid-model approximations, which are akin to our strong-law approximation, in their work.

ACKNOWLEDGMENTS

The first author was supported by a National Defense Science and Engineering Graduate Fellowship. This work was partially supported by National Science Foundation grant DMI 0400287.

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.

Billingsley, P. 1968. *Convergence of Probability Measures*. New York: Wiley.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspec-

tive *Journal of the American Statistical Association* 100:36–50

Buist, E , and P L'Ecuyer 2005 A java library for simulating contact centers In *Proceedings of the 2005 Winter Simulation Conference*, ed M E Kuhl, N M Steiger, F B Armstrong, and J A Joines Piscataway NJ: IEEE

Chen, B P K , and S G Henderson 2001 Two issues in setting call center staffing levels *Annals of Operations Research* 108:175–192

Cleveland, B , and J Mayben 1997 *Call Center Management on Fast Forward* Annapolis, MD: Call Center Press

Green, L V , P J Kolesar, and J Soares 2001 Improving the SIPP approach for staffing service systems that have cyclic demands *Operations Research* 49 (4): 549–564

Harrison, J M , and A Zeevi 2005 A method for staffing large call centers using stochastic fluid models *Manufacturing & Service Operations Management* To appear

Ittig, P 1994 Planning service capacity when demand is sensitive to delay *Decision Sciences* 25 (4): 541–559

Ittig, P 2002 The real costs of making customers wait *International Journal of Service Industry Management* 13 (3): 231–241

Mehrotra, V 1997 Ringing up big business *OR/MS Today* 24 (4): 18–24

Meyn, S P , and R L Tweedie 1993 *Markov Chains and Stochastic Stability* London: Springer-Verlag

Pullman, M , and W Moore 1999 Optimal service design: integrating marketing and operations perspectives *International Journal of Service Industry Management* 10 (2): 239–260

Ross, S M 1996 *Stochastic Processes* 2nd ed New York: Wiley

Steckley, S G , S G Henderson, and V Mehrotra 2005 Service system planning in the presence of a random arrival rate Working paper

Whitt, W 1976 Bivariate distributions with given marginals *The Annals of Statistics* 4:1280–1289

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

Whitt, W 2004 Staffing a call center with uncertain arrival rate and absenteeism Submitted

AUTHOR BIOGRAPHIES

SAM G. STECKLEY is a Ph D candidate in the School of Operations Research and Industrial Engineering at Cornell University His primary field of interest is input model uncertainty in discrete-event simulation He is the recipient of an NDSEG fellowship His e-mail address is ⟨steckley@orie.cornell.edu⟩

SHANE G. HENDERSON is an associate professor in the School of Operations Research and Industrial Engineering at Cornell University He has previously held positions in the Department of Industrial and Operations Engineering at the University of Michigan and the Department of Engineering Science at the University of Auckland He is an associate editor for the *ACM Transactions on Modeling and Computer Simulation*, *Operations Research Letters*, and *Mathematics of Operations Research*, and the secretary of the INFORMS Simulation Society He likes cats but is allergic to them His research interests include discrete-event simulation and simulation optimization His e-mail address is ⟨sgh9@cornell.edu⟩, and his web page is ⟨www.orie.cornell.edu/~shane⟩

VIJAY MEHROTRA is a faculty member in the Department of Decision Sciences at San Francisco State University and a management consultant with a specialization in call center operations He holds a Ph D in Operations Research from Stanford University and a B A from St Olaf College His column "Was It Something I Said?" has appeared in **OR/MS Today** since 1997 He can be reached via email at ⟨vjm@sfsu.edu⟩ or via his web page at ⟨online.sfsu.edu/~drvijay⟩