OPERATIONS RESEARCH
Vol. 61, No. 4, July–August 2013, pp. 1030–1045
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2013.1193

Statistical Analysis with Little's Law

Song-Hee Kim, Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
{sk3116@columbia.edu, ww2040@columbia.edu}

The theory supporting Little's Law ($L = \lambda W$) is now well developed, applying to both limits of averages and expected values of stationary distributions, but applications of Little's Law with actual system data involve measurements over a finite-time interval, which are neither of these. We advocate taking a statistical approach with such measurements. We investigate how estimates of L and λ can be used to estimate W when the waiting times are not observed. We advocate estimating confidence intervals. Given a single sample-path segment, we suggest estimating confidence intervals using the method of batch means, as is often done in stochastic simulation output analysis. We show how to estimate and remove bias due to interval edge effects when the system does not begin and end empty. We illustrate the methods with data from a call center and simulation experiments.

Subject classifications: Little's Law; $L = \lambda W$; measurements; parameter estimation; confidence intervals; bias; finite-time versions of Little's Law; confidence intervals with $L = \lambda W$; edge effects in $L = \lambda W$; performance analysis.
Area of review: Stochastic Models.
History: Received March 2012; revisions received August 2012, December 2012; accepted May 2013. Published online in *Articles in Advance* August 7, 2013.

1. Introduction

We have just celebrated the 50th anniversary of the famous paper by Little (1961) on the fundamental queueing relation $L = \lambda W$ with a retrospective by Little (2011) himself, which emphasizes the applied relevance as well as reviewing the advances in theory, including the sample-path proof by Stidham (1974) and the extension to $H = \lambda G$. Several books provide thorough treatments of the theory, including the sample-path analysis by El-Taha and Stidham (1999) and the stationary framework involving the Palm transformation by Sigman (1995) and Baccelli and Bremaud (2003), as well as the perspective within stochastic networks by Serfozo (1999). As a consequence, $L = \lambda W$ and the related conservation laws are now on a solid mathematical foundation.

The relation $L = \lambda W$ can be quickly stated: The average number of customers waiting in line (or items in a system), L, is equal to the arrival rate (or throughput) λ multiplied by the average waiting time (time spent in system) per customer, W. If we know any two of these quantities, then we necessarily know all three. The easily understood reason is reviewed in §2. With queueing models where λ is known, the relation $L = \lambda W$ yields the value of L or W whenever the other has been calculated.

1.1. Measurements Over a Finite Time Interval

However, in many applications, these conservation laws are applied with measurements over a finite-time interval of length t, yielding finite averages $\bar{L}(t)$, $\bar{\lambda}(t)$, and $\overline{W}(t)$ (defined in (1) below). Indeed, the applied relevance with measurements motivated Little (2011) to discuss relations among finite-time measurements instead of the stationary framework in Little (1961). However, with finite averages, the large body of supporting theory often does not apply directly, because that theory concerns either long-run averages (limits) or the expected values of stationary stochastic processes in stochastic models. The available measurements are neither of these.

Here is the essence of a typical application: We start with the observation of $L(s)$, the number of items in the system at time s, for $0 \leqslant s \leqslant t$. From that sample path, we can directly observe the arrivals (jumps up) and departures (jumps down). Hence, we can easily estimate the arrival rate λ and the average number in system L. However, based only on the available information, we typically cannot determine the time each item spends in the system, because the items need not depart in the same order that they arrived. Nevertheless, we can estimate the average waiting time by $W = L/\lambda$, using our estimates of L and λ.

In this paper we focus on the typical application in the preceding paragraph, estimating W given estimates of L and λ, illustrated by data from a large call center. The first issue is that, with commonly accepted definitions (see (1) below), the relation $L = \lambda W$ is not valid as an equality over a finite-time interval unless the system starts and ends empty, which often is either not feasible or not desirable. In §2 we review the exact relation that holds for finite-time intervals and a way to modify the definitions so that the edge effects do not occur, even when the system does not start and end empty. Using modified definitions to make $\bar{L}(t) = \bar{\lambda}(t)\overline{W}(t)$ valid for *all* finite intervals is the approach of the "operational analysis" proposed by

Buzen (1976) and Denning and Buzen (1978), motivated by performance analysis of computer systems, which is also discussed by Little (2011). Changing definitions in that way can be very helpful to check the consistency of measurements and data analysis, which is a legitimate concern. Although changing the definitions is one option, we advocate *not* doing so, because it leads to problems with interpretation.

1.2. A Statistical Approach

We advocate taking a statistical approach with data over a finite-time interval. Thus we regard the finite averages as realizations of random estimators of underlying unknown "true" values. We suggest estimating confidence intervals. Because the initial estimators may be biased, we suggest refined estimators to reduce the bias. To the best of our knowledge, a statistical approach has not been taken previously in the literature on applications of $L = \lambda W$ with measurements; e.g., see Denning and Buzen (1978), Little and Graves (2008), Little (2011), Lovejoy and Desmond (2011) and Mandelbaum (2011).

1.2.1. A Stationary Framework. Two very different settings can arise: stationary and nonstationary. Preliminary data analysis should be done to determine if the data are from a stationary environment. In a stationary framework, we assume that Little's Law theory applies, so that L, λ, and W are well defined, corresponding to both means of stationary probability distributions and limits of averages (assumed to exist), and related by $L = \lambda W$. We thus regard the underlying parameters L, λ, and W as the *true* values that we want to estimate; we regard the averages $\bar{L}(t)$, $\bar{\lambda}(t)$, and $\overline{W}(t)$ based on measurements over a time interval $[0, t]$ as estimates of these parameters.

To learn how well we know L, λ, and W when we compute the averages $\bar{L}(t)$, $\bar{\lambda}(t)$, and $\overline{W}(t)$, we suggest estimating confidence intervals. Given a single sample path from an interval that can be regarded as approximately stationary, we suggest applying the method of batch means to estimate confidence intervals, as is commonly done in simulation output analysis, and has been studied extensively; e.g., see Alexopoulos and Goldsman (2004), Asmussen and Glynn (2007), Tafazzoli et al. (2011), Tafazzoli and Wilson (2011) and references therein. We present theory supporting its application in the present context.

In addition, we are concerned with the statistical problem of how to make inferences from limited data. We illustrate by focusing on estimating W given the finite averages $\bar{L}(t)$ and $\bar{\lambda}(t)$ when the waiting times are not directly observed. We pay special attention to the indirect estimator $\overline{W}_{L,\lambda}(t) \equiv \bar{L}(t)/\bar{\lambda}(t)$ suggested by Little's Law. We show the special definition used to obtain equality for $\bar{L}(t) = \bar{\lambda}(t)\overline{W}(t)$ within each subinterval seriously distorts the batch-means estimators when the modified definition is used within each subinterval.

1.2.2. A Nonstationary Framework. However, many applications with data involve nonstationary settings; e.g., service systems typically have arrival rates that vary significantly over each day. Estimation is more complicated without stationarity, because conventional Little's Law theory no longer applies. Indeed, the parameters L, λ, and W are typically no longer well defined. To specify what we are trying to estimate, we assume that there is an unspecified underlying stochastic queueing model, which may be highly nonstationary (for which the processes in §2.1 are well defined). As usual with Little's Law, it is not necessary to define the underlying queueing model in detail. Then we regard the vector of time averages $(\bar{L}(t), \bar{\lambda}(t), \overline{W}(t))$ as a random vector with an associated vector of finite mean values $(E[\bar{L}(t)], E[\bar{\lambda}(t)], E[\overline{W}(t)])$. We propose that mean vector as the quantity to be estimated.

Since the method of batch means is no longer appropriate without stationarity, we suggest an approach corresponding to independent replications. That approach is appropriate for call centers when the data comes from multiple days that can be regarded as independent and identically distributed. In a nonstationary setting, the bias can be much more important, so we discuss ways to reduce it.

1.2.3. Validation by Simulation. Because actual system data may be complicated and limited, we suggest applying simulation to study how the estimation procedures proposed here work for an idealized queueing model of the system. In doing so, we presume that we do not know enough about the actual system to construct a model that we can directly apply to compute what we are trying to estimate, but that we know enough to be able to construct an idealized model to evaluate how the proposed estimation procedures perform. We illustrate this suggested simulation approach with our call center example in §3.2.

1.3. Organization

Here is how the rest of this paper is organized: In §2 we discuss the finite-time version of $L = \lambda W$, emphasizing the interval edge effects. In §3 we apply the statistical approach to a banking call center example and associated simulation models. In §4 we study ways to estimate confidence intervals. In §5 we study ways to estimate and reduce the bias in the estimator $\overline{W}_{L,\lambda}(t) \equiv \bar{L}(t)/\bar{\lambda}(t)$. In §6 we perform experiments combining the insights in §§4 and 5; we estimate confidence intervals for refined estimators designed to reduce bias. Finally, in §7 we draw conclusions. Additional material appears in the e-companion and a technical report (Kim and Whitt 2012) is available on the authors' web pages; the contents of both are described at the beginning of the e-companion. Kim and Whitt (2013) is a sequel to this paper on estimating waiting times with the time-varying Little's Law. Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2013.1193.

2. Measurement Over a Finite Time Interval: Definitions and Relations

In this section we review analogs of $L = \lambda W$ for a finite-time interval, denoted by $[0, t]$. Consistent with most applications, we assume that the system was in operation in the past, prior to time 0, and that it will remain in operation after time t. We will use standard queueing terminology, referring to the items being counted as customers. We focus on the customers that are in the system at some time during the interval $[0, t]$. Let these customers be indexed in order of their arrival time, which could be prior to time 0 if the system is not initially empty (with some arbitrary method to break ties, if any).

2.1. The Performance Functions and Their Averages

For customer k, let A_k be the arrival time, D_k the departure time, and $W_k \equiv D_k - A_k$ the waiting time (time in system), where $-\infty < A_k < D_k < \infty$, $[0, t] \cap [A_k, D_k] \neq \varnothing$, and $\equiv$ denotes "equality by definition." Let $R(0)$ count the customers that arrived before time 0 that remain in the system at time 0; let $A(t)$ count the total number of new arrivals in the interval $[0, t]$; and let $L(t)$ be the number of customers in the system at time t. Thus, $A(t) = \max\{k \geqslant 0\colon A_k \leqslant t\} - R(0)$, $t \geqslant 0$, and $L(0) = R(0) + A(0)$, where $A(0)$ is the number of new arrivals at time 0, if any. We will carefully distinguish between $R(0)$ and $L(0)$, but the common case is to have $A(0) = 0$ and $L(0) = R(0)$.

The respective averages over the time interval $[0, t]$ are

$$\bar{\lambda}(t) \equiv t^{-1}A(t), \quad \bar{L}(t) \equiv t^{-1}\int_0^t L(s)\,ds,$$
$$\overline{W}(t) \equiv (1/A(t))\sum_{k=R(0)+1}^{R(0)+A(t)} W_k, \tag{1}$$

where $0/0 \equiv 0$ for $\overline{W}(t)$. The first two are time averages, whereas the last, $\overline{W}(t)$, is a customer average, but over all arrivals during the interval $[0, t]$.

We will focus on these averages over $[0, t]$ in (1), but we could equally well consider the averages associated with the first n arrivals. To do so, let T_n be the arrival epoch of the nth new arrival, i.e., $T_n \equiv A_{n+R(0)}$, $n \geqslant 0$,

$$\bar{\lambda}_n \equiv n/T_n, \quad \bar{L}_n \equiv (1/T_n)\int_0^{T_n} L(s)\,ds,$$
$$\overline{W}_n \equiv n^{-1}\sum_{k=R(0)+1}^{R(0)+n} W_k. \tag{2}$$

As in (1), the first two averages in (2) are time averages, but over the time interval $[0, T_n]$, whereas the last, $\overline{W}_n$, is a customer average over the first n (new) arrivals. If there is only a single arrival at time T_n, then the averages in (2) can be expressed directly in terms of the averages in (1): $\bar{\lambda}_n = \bar{\lambda}(T_n)$, $\bar{L}_n = \bar{L}(T_n)$, and $\overline{W}_n = \overline{W}(T_n)$, so that conclusions for (1) yield analogs for (2).

Just as we can use the relation $L = \lambda W$ and knowledge of any two of the three quantities L, λ, and W to compute the remaining one, so we can use any two of the three estimators in (1) to create a new alternative estimator, exploiting $L = \lambda W$:

$$\bar{L}_{W,\lambda}(t) \equiv \bar{\lambda}(t)\overline{W}(t), \quad \bar{\lambda}_{L,W}(t) \equiv \frac{\bar{L}(t)}{\overline{W}(t)}, \quad \text{and}$$
$$\overline{W}_{L,\lambda}(t) \equiv \frac{\bar{L}(t)}{\bar{\lambda}(t)}. \tag{3}$$

For the typical application mentioned in §1 in which we observe $L(s)$, $0 \leqslant s \leqslant t$, we can directly construct the averages $\bar{L}(t)$ and $\bar{\lambda}(t)$, but we may not observe the individual waiting times. Hence, we may want to use $\overline{W}_{L,\lambda}(t)$ in (3) as a substitute for $\overline{W}(t)$ in (1).

2.2. How the Averages in (1) Are Related

Figures 1 and 2 show how the three averages in (1) are related. These averages are related via $\bar{L}(t) = \bar{\lambda}(t)\overline{W}(t)$ if the system starts and ends empty, i.e., if $R(0) = L(t) = 0$,

Figure 1. The total work in the system during the interval $[0, t]$ with edge effects: Including arrivals before time 0 and departures after time t.

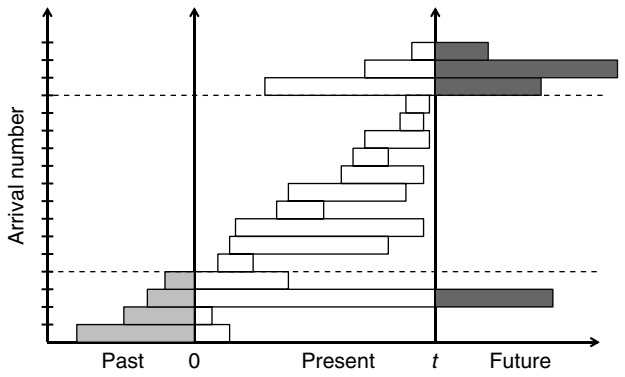
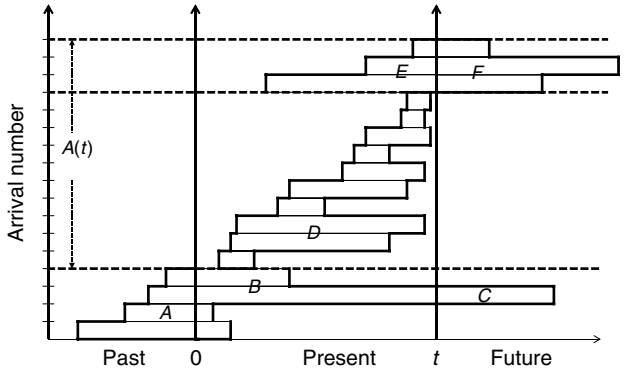


Figure 2. Six regions: Waiting times (i) of customers that both arrive and depart inside $[0, t]$ (D), (ii) of arrivals before time 0 ($A \cup B \cup C$), and (iii) of departures after time t ($C \cup E \cup F$).



as we show in Theorem 1. However, more generally, these averages are not simply related. To illustrate, in Figures 1 and 2 a bar of height 1 is included for each of the customers in the system at some time during $[0, t]$ with the bar extending from the customer's arrival time to its departure time. (In this example the customers do not depart in the same order they arrived.) Thus the width of the bar is the customer's waiting time. For $0 \leqslant s \leqslant t$, the number of bars above any time s is $L(s)$.

To better communicate what is going on visually, we have ordered the customers in a special way. In Figures 1 and 2, the customers that arrive before time 0 but are still there at time 0 are placed first, starting at the bottom and proceeding upwards. These customers are ordered according to the arrival time, so the customers that arrived before time 0 appear at the bottom. One of these customers also departs after time t. The customers that arrived before time 0 and are still in the system at time 0 contribute to the regions A, B, and C in Figure 2.

After the customers that arrived before time 0, we place the customers that arrive after time 0 and depart before time t, in order of arrival; they constitute region D in Figure 2. Finally, we place the customers that arrive after time 0 but depart after time t. These customers are ordered according to their arrival time as well; they constitute regions E and F in Figure 2. Three extra horizontal lines are included, along with the vertical lines at times 0 and t, to separate the regions. The arrival numbers are indicated along the vertical y axis. The condition $R(0) = L(t) = 0$ arises in Figure 2 as the special case in which all regions except region D are empty.

The averages can be expressed in terms of the two *cumulative processes*,

$$\begin{aligned} &C_L(t) \equiv \int_0^t L(s)\,ds \quad \text{and} \\ &C_W(t) \equiv \sum_{k=R(0)+1}^{R(0)+A(t)} W_k, \quad t \geqslant 0. \end{aligned} \tag{4}$$

The difference between these two cumulative processes can be expressed in terms of the process $T_W^{(r)}(t)$, recording the *total residual waiting time* of all customers in the system at time t, i.e.,

$$T_W^{(r)}(t) \equiv \sum_{k=1}^{L(t)} W_k^{r,\,t}, \tag{5}$$

where $W_k^{r,\,t}$ is the remaining waiting time at time t for customer k in the system at time t (with index k assigned at time t among those remaining). The averages in (1) are the *time average* $\bar{L}(t) \equiv t^{-1}C_L(t)$ and the *customer average* $\overline{W}(t) \equiv C_W(t)/A(t)$. For a region A in Figure 2, let $|A|$ be the area of A. In general, the cumulative processes can be expressed in terms of the regions in Figure 2 as $C_L(t) = |B \cup D \cup E|$ and $C_W(t) = |D \cup E \cup F|$, whereas $T_W^{(r)}(0) = |B \cup C|$ and $T_W^{(r)}(t) = |C \cup F|$, so that

$$\begin{aligned} C_L(t) - C_W(t) &= |B| - |F| = |B \cup C| - |F \cup C| \\ &= T_W^{(r)}(0) - T_W^{(r)}(t). \end{aligned} \tag{6}$$

This relation for $C_L(t)$ is easy to see if we let ν be the total number of arrivals and departures in the interval $[0, t]$, τ_k be the kth-ordered time point among all the arrival times and departure times in $[0, t]$, with ties indexed arbitrarily and consistently, $\tau_0 \equiv 0$ and $\tau_{\nu+1} = t$. Then

$$\begin{aligned} C_L(t) &\equiv \int_0^t L(s)\,ds = \sum_{j=1}^{\nu+1} \int_{\tau_{j-1}}^{\tau_j} L(s)\,ds \\ &= \sum_{j=1}^{\nu+1} L(\tau_{j-1})(\tau_j - \tau_{j-1}) = |B \cup D \cup E|, \end{aligned}$$

where the last relation holds because $L(\tau_{j-1})$ is the number of single-customer unit-height bars above the interval $[\tau_{j-1}, \tau_j]$. Since $C_L(t) = C_W(t) = |D|$ if $R(0) = L(t) = 0$, we necessarily have the following well-known result, appearing as Theorem I of Jewell (1967).

THEOREM 1 (TRADITIONAL FINITE-TIME LITTLE'S LAW). *If* $R(0) = L(t) = 0$, *then* $\bar{L}(t) = \bar{\lambda}(t)\overline{W}(t)$.

PROOF. Under the condition, $\bar{L}(t) \equiv C_L(t)/t = C_W(t)/t = (A(t)/t)(C_W(t)/A(t)) \equiv \bar{\lambda}(t)\overline{W}(t)$. □

On the other hand, for the common case in which there are customers in the system during $[0, t]$ that arrived before time 0 and/or depart after time t, as in Figures 1 and 2, there is no simple relation between these cumulative processes and the associated averages, because of the interval edge effects. Nevertheless, the analysis above exposes the relationship that does hold. Variants of these relations are needed to establish sample-path limits in Little law theory, so the following result should not be considered new; e.g., see Glynn and Whitt (1986, Theorem 1). A variant appears in Mandelbaum (2011, p. 17.4), who credits it to his student Abir Koren and emphasizes its importance for looking at data.

THEOREM 2 (EXTENDED FINITE-TIME LITTLE'S LAW). *The averages in* (1) *and* (3) *are related by*

$$\begin{aligned} \Delta_L(t) &\equiv \bar{L}_{W,\,\lambda}(t) - \bar{L}(t) = \frac{|F| - |B|}{t} \\ &= \frac{T_W^{(r)}(t) - T_W^{(r)}(0)}{t}; \\ \Delta_W(t) &\equiv \overline{W}_{L,\,\lambda}(t) - \overline{W}(t) = \frac{|B| - |F|}{A(t)} = -\frac{\Delta_L(t)}{\bar{\lambda}(t)} \\ &= \frac{T_W^{(r)}(0) - T_W^{(r)}(t)}{A(t)}; \\ \Delta_\lambda(t) &\equiv \bar{\lambda}_{L,\,W}(t) - \bar{\lambda}(t) = \left(\frac{|B| - |F|}{|D| + |E| + |F|}\right)\bar{\lambda}(t) \\ &= -\frac{\Delta_L(t)}{\overline{W}(t)}, \end{aligned} \tag{7}$$

where $|B|$ is the area of the region B in Figure 2 and $T_W^{(r)}(t)$ is defined in (5).

Since we focus on inferences about the average wait based on $\bar{L}(t)$ and $\bar{\lambda}(t)$ using $\overline{W}_{L,\lambda}(t)$, we focus on $\Delta_W(t)$ in (7). Given that the customers need not depart in the order they arrive and we only observe $L(s)$, $0 \leqslant s \leqslant t$, the random variables $T_W^{(r)}(0)$ and $T_W^{(r)}(t)$ appearing in $\Delta_W(t)$ in (7) are not directly observable; we only have partial information about these random variables.

2.3. Alternative Definitions to Force Equality: The Inside View

Denning and Buzen (1978), Little (2011), and others have observed that we can preserve the relation $\bar{L}(t) = \bar{\lambda}(t)\overline{W}(t)$ in Theorem 1 without any conditions on $R(0)$ and $L(t)$ if we change the definitions. Equality can be achieved in general if we assume that our entire view of the system is *inside* the interval $[0, t]$. We see arrivals before time 0 but only as arrivals appearing at time 0, and we see the portions of all waiting times only within the interval $[0, t]$. To achieve the inside view, let $A^{(i)}(t)$ count the number of new arrivals *plus* the number of customers initially in the system, and let $W_k^{(i)}$ measure the waiting time *inside* the interval $[0, t]$; i.e., let

$$\begin{aligned} &A^{(i)}(t) \equiv R(0) + A(t), \quad t \geqslant 0, \quad \text{and} \\ &W_k^{(i)} \equiv (D_k \wedge t) - (A_k \vee 0), \quad k \geqslant 1, \end{aligned} \tag{8}$$

where $a \wedge b \equiv \min\{a, b\}$ and $a \vee b \equiv \max\{a, b\}$. Now consider the associated averages

$$\bar{\lambda}^{(i)}(t) \equiv t^{-1}A^{(i)}(t) \quad \text{and} \quad \overline{W}^{(i)}(t) \equiv \frac{\sum_{k=1}^{A^{(i)}(t)} W_k^{(i)}}{A^{(i)}(t)}. \tag{9}$$

By an elementary modification of the proof of Theorem 1, we obtain the following "operational analysis" relation. (The equality relation corresponds to the operational Little's Law of Denning and Buzen 1978, p. 235; Little 2011, Theorem LL.2.)

THEOREM 3 (FINITE-TIME VERSION OF LITTLE'S LAW WITH ALTERED DEFINITIONS). *With the new definitions in (8) and (9), $\bar{\lambda}^{(i)}(t) \geqslant \bar{\lambda}(t)$, $\overline{W}^{(i)}(t) \leqslant \overline{W}(t)$, and $\bar{L}(t) = \bar{\lambda}^{(i)}(t)\overline{W}^{(i)}(t)$.*

Given that we only see inside the interval $[0, t]$, the reduced waiting times are *censored*. Indeed, there is no valid upper bound on $\overline{W}(t)$ based on the inside view. Arrivals before time 0 can have occurred arbitrarily far in the past prior to time 0, and customers present at time t can remain arbitrarily far into the future after time t. Any further properties of $\overline{W}(t)$ must depend on additional assumptions about what happens outside the interval $[0, t]$.

Even though the new definitions provide a good framework for checking the consistency of the data processing, and can be regarded as proper definitions, we advocate *not* using this modification because it causes problems in interpretation. We think it is usually better to account for the fact that an important part of the story takes place *outside* the interval $[0, t]$, even if we do not see it all. The alternative definitions in (8) also cause problems with the method of batch means used to construct confidence intervals; see §3.4.

3. A Banking Call Center Example

We illustrate the statistical approach by considering data from a telephone call center of an American bank from the data archive of Mandelbaum (2012). In 2001, this banking call center had sites in four states, which were integrated to form a single virtual call center. The virtual call center had 900–1,200 agent positions on weekdays and 200–500 agent positions on weekends. The center processed about 300,000 calls per day during weekdays, with about 60,000 (20%) handled by agents, with the rest being served by integrated voice response (IVR) technology. As in many modern call centers, in this banking call center there were multiple agent types and multiple call types, with a form of skill-based routing (SBR) used to assign calls to agents.

Because we are only concerned with estimation related to the three parameters L, λ, and W, we do not get involved with the full complexity of this system. For this paper, we use data for one class of customers, denoted by "Summit," for 18 weekdays in May 2001; the data used and the analysis procedure are available from the authors' Websites. Each working day covers a 17-hour period from 6 A.M. to 11 P.M., referred to as $[6, 23]$.

3.1. Sample Paths for a Typical Day

For some of the analyses, we will use a single day, Friday, May 25, 2001. Over this 17-hour period on that one day there were 5,749 call arrivals (of this one type requesting an agent), of which 253 (4.4%) abandoned from queue before starting service. We do not include these abandonments in our analysis. Figures 3 and 4 show plots of the total number of arrivals into the queue (system), $A_q(s)$, and into service, $A_{\text{ser}}(s)$, together with the total number of departures from the queue (system), $D_q(s)$, and from service, $D_{\text{ser}}(s)$, all over the interval $[0, s]$, $0 \leqslant s \leqslant t$, first over the entire working day $[6, 23]$ and then over the hour $[14, 15]$. These are based on the counts over one-second subintervals. Note that the four curves in Figure 3 are too close to discern due to short waiting time (time in system, measured in minutes) relative to the time scale (hours). We see better when we zoom in, as in Figure 4.

From the first plot in Figure 3, we see that the arrival rate is *not* stationary over the entire day (because the slope is not nearly constant), but it appears to be approximately stationary over the middle part of the day, e.g., in the six-hour interval $[10, 16]$. When the arrival rate is nearly constant, so is the departure rate. The stationary-and-independent-increments property associated with a

Figure 3. Arrivals and departures over the full day of May 25, 2001.

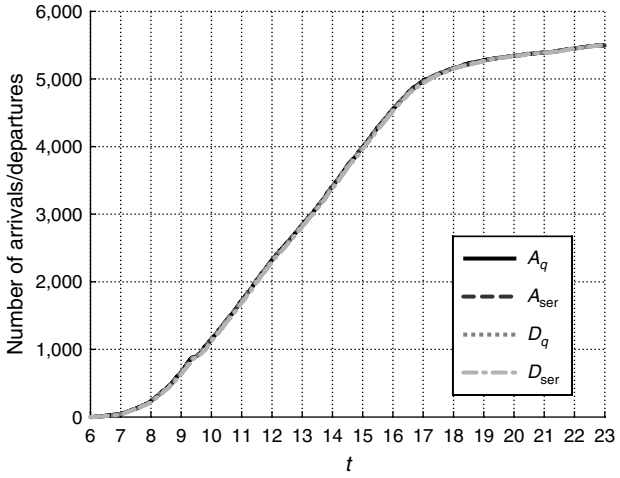
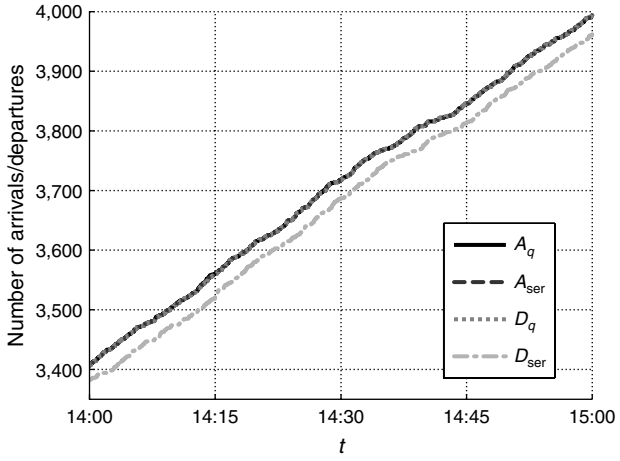


Figure 4. Arrivals and departures over the hour [14, 15] within that day.



homogeneous Poisson process over [10, 16] and the nonstationarity over [6, 10] and [16, 23] were confirmed by applying the turning points test, the difference-sign test and the rank test for randomness discussed in Brockwell and Davis (1991, p. 312); the details appear in §2 of the e-companion.

To confirm what we deduce from the arrival and departure rates, we also plot the number in system L_{sys} and the waiting times (times spent in the system), W_{sys}, and their hourly averages over the full day in Figures 5 and 6. Consistent with the plots in Figure 3, we see that the number in system looks approximately stationary in the 6-hour interval [10, 16], but not over the full day [6, 23]. In addition, Figure 6 shows that the hourly averages of the waiting times do not change much, especially in the interval [10, 16]. During that six-hour period [10, 16], during which the system is approximately stationary, agents handled 3,427 calls, of which only 28 (0.82%) abandoned. However, closer examination shows that the sample means $\bar{L}$ are 28.3 and 32.6 over the hours [13, 14] and [14, 15],

Figure 5. L_{sys} and its hourly averages over the full day.

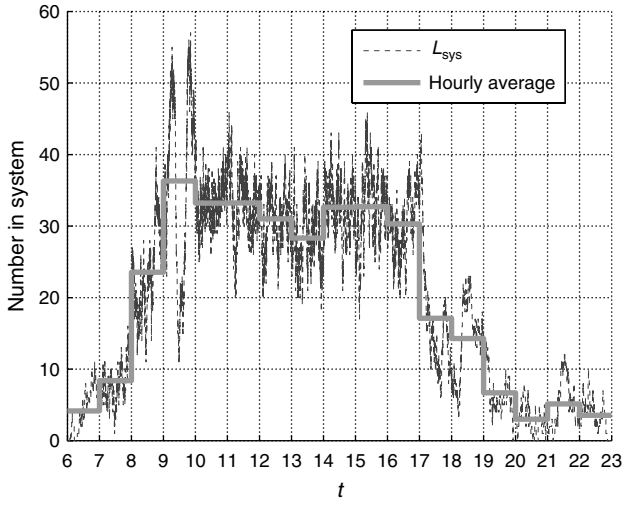
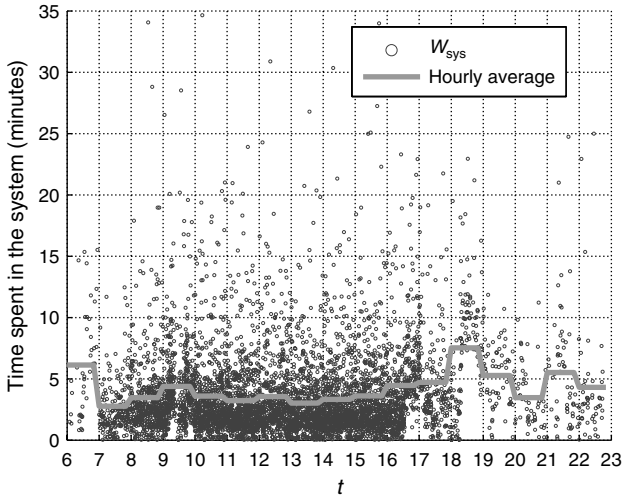


Figure 6. W_{sys} and its hourly averages over the full day.



respectively, so that the differences can be shown to be statistically significant, but are minor compared to differences at the ends of the day. Since stationarity clearly does not hold exactly, caution should be used with the estimates.

To illustrate both the statistical approach to this example and the consequence of nonstationarity, we estimated L, λ, and W both over the full day [6, 23] and over the approximately stationary subinterval [10, 16]. For both, we used the method of batch means, dividing the interval into $m = 20$ batches of equal length, producing batch lengths of 51 and 18 minutes, respectively. Over the full day, we have the estimates (measuring time in minutes)

$$\bar{L}_{full} = 20.2 \pm 6.1, \quad \bar{\lambda}_{full} = 5.39 \pm 1.84, \quad \text{and} \quad \bar{W}_{full} = 4.18 \pm 0.56; \tag{10}$$

over the interval [10, 16], we have the estimates

$$\bar{L}_{stat} = 31.8 \pm 1.0, \quad \bar{\lambda}_{stat} = 9.44 \pm 0.31, \quad \text{and} \quad \bar{W}_{stat} = 3.39 \pm 0.15. \tag{11}$$

For each estimate in (10) and (11), we also include the halfwidth of the 95% confidence interval, estimated as described in §4.3. We draw the following conclusions: (i) the confidence intervals tell us more than the averages alone, (ii) paying attention to stationarity is important, (iii) the halfwidths themselves reveal the nonstationarity, because we get far smaller halfwidths with the shorter subinterval $[10, 16]$, and (iv) since the mean waiting time is much less than the batch length, the number of batches is not grossly excessive (but that requires further examination).

3.2. Supporting Call Center Simulation Models

Many-server systems such as call centers are characterized by having many servers working independently in parallel. In such systems (if managed properly), the waiting times in queue tend to be short compared to the service times, and the service times tend to be approximately i.i.d. and independent of the arrival process. Thus, it is natural to use an idealized *infinite-server paradigm*, involving an infinite-server (IS) model with i.i.d. service times independent of the arrival process to approximately analyze statistical methods. Because the service times coincide exactly with the waiting times in the IS model, the waiting times are i.i.d. with constant mean $E[S]$, even though we are considering a nonstationary setting. That often holds approximately in service systems, as illustrated by our call center example.

For the call center, we have data on the arrival times and waiting times as well as the number in system $L(s)$, $0 \leqslant s \leqslant t$, but we do not have data on the staffing and the complex call routing. Thus, as suggested in §1.2.2, to evaluate the estimation procedures, we simulate the single-class, single-service-pool $M_t/GI/\infty$ IS model and associated $M_t/GI/s_t$ models with time-varying staffing levels chosen to yield good performance, exploiting the square root staffing (SRS) formula $s(t) \equiv m(t) + \beta\sqrt{m(t)}$, where $m(t)$ is the *offered load*, the time-varying mean number of busy servers in the IS model, as in Jennings et al. (1996). As described in §3.1 of the e-companion, we fit the arrival rate function to a continuous piecewise-linear function, with one increasing piece over $[6, 10]$ starting at 0, a constant piece over $[10, 16]$, and two decreasing linear pieces over $[16, 18]$ and $[18, 23]$, the first steeper and the second ending at 0. We then simulated a nonhomogeneous Poisson arrival process with this arrival rate function. We assumed that all the service times were i.i.d., with a distribution obtained to match the observed waiting time distribution. A lognormal distribution with mean 3.38 and squared coefficient of variation $c_s^2 = 1.02$ was found to be a good fit, but an exponential distribution with that mean (and $c_s^2 = 1$) was also a good approximation, and so was used, because it is easier to analyze (see §3.2 of the e-companion). The IS model was simulated with that fitted arrival rate function and service-time distribution. The offered load $m(t)$ was also computed by formulas (6) and (7) of Jennings et al. (1996), drawing on Eick et al. (1993b); then the staffing function $s(t)$ was determined by the SRS formula using a range of quality-of-service (QoS) parameters β (see §3.3 of the e-companion). We simulated 1,000 independent replications of each of these models to study how the methods to estimate confidence intervals performed. In the next subsection, we report results from simulation experiments showing that the finite-server models perform much like the IS model.

3.3. Confidence Intervals for the Call Center Data and Simulation

We applied the method of batch means to estimate confidence intervals for the parameters L, λ, and W using the direct sample averages from (1) plus indirect estimate $\overline{W}_{L,\lambda}(t)$ from (3) for the time interval $[10, 16]$ over which the system is approximately stationary. (For both the call center data and the simulation model, we observe the waiting times, but we examine the alternative estimator $\overline{W}_{L,\lambda}(t)$ from (3) to see how it would perform if we could not observe the waiting times.)

We also consider the idealized $M_t/M/\infty$ and $M_t/M/s_t$ simulation models introduced in §3.2 and explained in detail in §3 of the e-companion. The estimation results are shown in Table 1. Additional results with more values of m appear in Kim and Whitt (2012, Tables 4–9).

For large QoS parameter β, e.g., $\beta \geqslant 2.0$, the performance in the finite-server model is essentially the same as in the associated IS model, as can be seen from Table 1. However, as β decreases, more customers have to wait before starting service. Thus, the estimated mean waiting time increases from 3.38 in the IS model to 3.39 and 3.44, respectively, for $\beta = 1.5$ and 1.0, respectively. Similarly, the estimated mean number in the system increases from 31.5 to 31.6 and 32.1 for these same cases. Of special interest is the confidence interval (CI) coverage in the simulations based on 1,000 replications. Table 1 shows it is excellent for all values of m, being very close to the target 95.0%, for all $\beta \geqslant 1.5$. However, we see a drop in coverage for $\beta = 1$. Thus, to be conservative, we advocate using for the call center model the largest estimated CI, which usually should be associated with the smallest number of batches $m = 5$ for the call center data. Overall, Table 1 shows that the indirect estimator $\overline{W}_{L,\lambda}^{(m)}(t)$ behaves very much the same as the direct estimator $\overline{W}(t)$. Indeed, that is consistent with the theory and other experiments in this paper.

To illustrate what happens with a shorter sample-path segment, we consider the interval $[14, 15]$. Table 2 shows the corresponding estimates for the IS model and the call center. Additional results with more values of m appear in Tables 10 and 11 of Kim and Whitt (2012). In this case, $m = 5$, 10, and 20 corresponds to 5 batches of 12 minutes, 10 batches of 6 minutes, and 20 batches of 3 minutes, respectively. The CI coverage is again excellent for the IS model for all cases. However, since the mean waiting time

Table 1. Direct estimates of L, λ, and W from (1) plus indirect estimate $\overline{W}_{L,\lambda}(t)$ from (3) with associated 95% confidence intervals for the approximately stationary time interval $[10, 16]$, constructed using batch means for $m = 5, 10$, and 20 batches for the call center data and idealized simulation models, including the $M_t/M/\infty$ and $M_t/M/s_t$ models with piecewise-linear arrival rate function fit to data, mean service time of 3.38 minutes and time-varying staffing based on the square-root-staffing formula using QoS parameter β taking values ranging from 1.0 to 2.5.

Case	m	$\bar{L}^{(m)}(t)$	$\bar{\lambda}^{(m)}(t)$	$\overline{W}^{(m)}(t)$	Cov. (%)	$\overline{W}^{(m)}_{L,\lambda}(t)$	Cov. (%)
$\beta = \infty$	5	31.5 ± 2.0	9.33 ± 0.42	3.38 ± 0.15	95.1	3.38 ± 0.15	95.4
$(M_t/M/\infty)$	10	31.5 ± 1.6	9.33 ± 0.35	3.38 ± 0.13	95.0	3.38 ± 0.13	95.7
	20	31.5 ± 1.4	9.33 ± 0.33	3.38 ± 0.12	94.4	3.38 ± 0.12	95.3
$\beta = 2.5$	5	31.5 ± 2.0	9.33 ± 0.42	3.38 ± 0.15	95.3	3.38 ± 0.15	95.9
$(M_t/M/s_t)$	10	31.5 ± 1.6	9.33 ± 0.35	3.38 ± 0.13	95.2	3.38 ± 0.13	95.8
	20	31.5 ± 1.4	9.33 ± 0.33	3.38 ± 0.12	95.0	3.38 ± 0.12	95.3
$\beta = 2.0$	5	31.5 ± 2.0	9.33 ± 0.42	3.38 ± 0.16	95.2	3.38 ± 0.16	95.7
	10	31.5 ± 1.6	9.33 ± 0.35	3.38 ± 0.13	95.3	3.38 ± 0.13	95.6
	20	31.5 ± 1.4	9.33 ± 0.33	3.38 ± 0.12	95.0	3.38 ± 0.12	95.5
$\beta = 1.5$	5	31.6 ± 2.2	9.33 ± 0.42	3.39 ± 0.17	95.8	3.39 ± 0.17	95.9
	10	31.6 ± 1.7	9.33 ± 0.35	3.39 ± 0.14	94.9	3.39 ± 0.14	95.1
	20	31.6 ± 1.5	9.33 ± 0.33	3.39 ± 0.13	94.0	3.40 ± 0.13	94.9
$\beta = 1.0$	5	32.1 ± 2.6	9.33 ± 0.42	3.44 ± 0.21	95.0	3.44 ± 0.21	95.3
	10	32.1 ± 2.1	9.33 ± 0.35	3.44 ± 0.17	93.2	3.44 ± 0.17	93.5
	20	32.1 ± 1.8	9.33 ± 0.33	3.44 ± 0.15	91.4	3.44 ± 0.15	92.5
Data	5	31.9 ± 1.9	9.44 ± 0.49	3.38 ± 0.22		3.38 ± 0.19	
(call center)	10	31.9 ± 1.3	9.44 ± 0.36	3.39 ± 0.15		3.38 ± 0.16	
	20	31.9 ± 1.0	9.44 ± 0.30	3.39 ± 0.15		3.38 ± 0.11	

Note. Estimated confidence interval coverage is shown for the two waiting-time estimates for the simulations based on 1,000 replications.

Table 2. Direct estimates of L, λ, and W from (1) plus indirect estimate $\overline{W}_{L,\lambda}(t)$ from (3) with associated 95% confidence intervals for the approximately stationary time interval $[14, 15]$ constructed using batch means for $m = 5, 10$, and 20 batches for the call center data and simulation of the idealized $M_t/M/\infty$ models, with piecewise-linear arrival rate function fit to data, mean service time of 3.38 minutes, and time-varying staffing based on the square-root-staffing formula using QoS parameter β.

Case	m	$\bar{L}(t)$	$\bar{\lambda}(t)$	$\overline{W}(t)$	Cov. (%)	$\overline{W}^{(m)}_{L,\lambda}(t)$	Cov. (%)
$\beta = \infty$	5	31.4 ± 4.0	9.32 ± 1.04	3.37 ± 0.37	95.6	3.38 ± 0.37	94.8
$(M_t/M/\infty)$	10	31.4 ± 2.9	9.32 ± 0.87	3.37 ± 0.32	95.8	3.40 ± 0.32	95.4
	20	31.4 ± 2.1	9.32 ± 0.82	3.37 ± 0.30	95.9	3.46 ± 0.32	94.3
Data	5	32.6 ± 1.9	9.82 ± 0.82	3.33 ± 0.21		3.33 ± 0.10	
(call center)	10	32.6 ± 1.6	9.82 ± 0.79	3.33 ± 0.21		3.34 ± 0.16	
	20	32.6 ± 1.3	9.82 ± 0.81	3.32 ± 0.23		3.43 ± 0.31	

Note. Estimated confidence interval coverage based on 1,000 replications is shown for the two waiting time estimates for the simulations.

is about 3.4 minutes, we regard only $m = 5$ appropriate for the interval $[14, 15]$. (We also note that the difference between $\overline{W}^{(m)}_{L,\lambda}(t)$ and $\overline{W}(t)$ in Table 2 becomes greater as m increases. See §3.4.1 for more discussion on this.)

3.4. Edge Effects and the Method of Batch Means

The issue of interval edge effects discussed in §2 becomes more serious with the method of batch means. For a fixed sample-path segment of length t and m batches, there are m intervals, each with edge effects, and each interval is of length t/m instead of t.

3.4.1. The Error Due to the Interval Edge Effects. Formula (7) shows that the difference between $\overline{W}_{L,\lambda}(t)$ and $\overline{W}(t)$ should be inversely proportional to t in a stationary setting, because the distribution of $T_W^{(r)}(t)$ is independent of t, whereas $\bar{\lambda}(t) \equiv t^{-1}A(t) \to \lambda$ as $t \to \infty$. We should expect serious bias if t is less than or equal to W, the average time spent in the system, but very little bias if t is much greater. Since $W \approx 3.4$ minutes for the call center example from §3, we expect serious bias if $t = 3$ minutes, some bias if $t = 30$ minutes and almost no bias if $t = 300$ minutes. Those expectations are confirmed by the averages shown in Table 3. In each case, the averages over subintervals correspond to batch means. (See Tables 12–14 of Kim and Whitt 2012 for more details.)

In Table 3 we see that the relative error $\Delta_W^{\mathrm{rel}}(t) \equiv \Delta_W(t)/\overline{W}_{L,\lambda}(t)$ takes the values 20.3%, 5.6%, and 0.5%,

Table 3. Comparison of the direct and indirect estimators $\overline{W}(t)$ and $\overline{W}_{L,\lambda}(t)$ for three values of t: 3, 30, and 300 minutes for the data from §3.

t	$\overline{W}(t)$	$\overline{W}_{L,\lambda}(t)$	$\lvert\Delta_W(t)\rvert$	$\Delta_W^{\text{rel}}(t)$ (%)	$\lvert U\rvert$	$\lvert A\rvert$ (%)	$\lvert B\rvert$ (%)	$\lvert C\rvert$ (%)	$\lvert D\rvert$ (%)	$\lvert E\rvert$ (%)	$\lvert F\rvert$ (%)	$\lvert F\rvert - \lvert B\rvert$ (%)
3	3.32	3.43	0.713	20.3	311	34.8	19.5	14.3	3.1	8.9	19.4	6.3
30	3.80	3.80	0.241	5.6	1,101	9.7	9.2	0.0	62.8	9.4	8.8	4.5
300	3.44	3.44	0.016	0.5	9,754	1.4	1.2	0.0	94.9	1.2	1.3	0.4

Note. Averages are given for the 20 subintervals of [14, 15] for $t = 3$ minutes, for the 20 subintervals of [8, 18] for $t = 30$ minutes and for the four overlapping five-hour subintervals of [9, 17], from [9, 14] to [12, 17], for $t = 300$ minutes.

respectively, for $t = 3$, 30, and 300 minutes. For the regions in Figure 2, for $t \geqslant 30$ minutes, we see that $|C| = 0$, the areas of regions B, C, E, and F are approximately independent of t, while the area of D is proportional to t. Table 3 shows the area of the union of all six regions, $U \equiv A \cup B \cup C \cup D \cup E \cup F$, and the percentages of that total area made up by each of the six regions, as well as $|F| - |B|$. The simple case occurs when region D dominates the six regions. The percentage of the total area provided by D is 94.9% for $t = 300$ minutes, 62.8% for $t = 30$ minutes, and 3.1% for $t = 3$ minutes.

3.4.2. Additional Error from the Altered Definitions. The altered definitions in §2.3 become more unattractive with batch means, because the shorter intervals distort the meaning even more. The average truncated waiting times $\overline{W}_c(t)$ in (9) tend to be even less than the true average waiting times W, whereas the average augmented arrivals $\bar{\lambda}_i(t)$ in (9) tend to be even more than the true average arrival rate λ. The altered definitions lead to double counting for arrivals. Customers that are in the system during more than one interval are counted as arrivals in all these intervals.

To illustrate, we consider the call center data over the interval [10, 16]. Without using batches, we have $\bar{\lambda}(t) = 9.44$ arrivals per minute and $\overline{W}(t) = 3.38$ minutes, whereas the estimators using the altered definitions in (9) are $\bar{\lambda}_i(t) = 9.55$ and $\overline{W}_c(t) = 3.33$. With m batches, $1 \leqslant m \leqslant 20$, the estimator $\bar{\lambda}(t)$ is unchanged and the estimator $\overline{W}(t)$ differs by only 0.001 from the original value of 3.38 for $m = 1$. In contrast, $\bar{\lambda}_i(t)$ assumes the values 9.55, 9.88, 10.33, and 11.16 for $m = 1$, 5, 10, and 20, respectively. Similarly, $\overline{W}_c(t)$ assumes the values 3.33, 3.22, 3.09, and 2.86 for $m = 1$, 5, 10, and 20, respectively. For $m = 20$, the errors in $\bar{\lambda}_i(t)$ and $\overline{W}_c(t)$ are 18% and 15%, respectively. When confidence intervals are formed based on batch means (for nonnegligible m), the systematic errors caused by the altered definition far exceed the halfwidth of the confidence interval. Hence, we recommend not using the modified definitions in (9).

4. Confidence Intervals: Theory and Methodology

We now consider how to apply the estimator $\overline{W}_{L,\lambda}(t)$ in (3) to estimate a confidence interval (CI) for W in a stationary setting and for $E[\overline{W}(t)]$ in a nonstationary setting, without observing the waiting times. We will be using statistical methods commonly used in simulation experiments. However, unlike simulation, we anticipate that system data is likely to be limited, so we may not be able to achieve high precision. Nevertheless, we want to have some idea how well we know the estimated values. With that in mind, we suggest applying standard statistical methods. To evaluate how well these statistical procedures should perform, e.g., to verify that CI coverage should be approximately as specified, we advocate studying associated idealized simulation models of the system more closely as suggested in §1.2.2 and as illustrated in §3.2.

For the common case in which we have only a single sample-path segment, we advocate applying the method of batch means, as specified in §4.3. That method depends on the batch means being approximately i.i.d. and normally distributed. We point out that there is a risk that these assumptions may not be justified, so that estimated CIs should be used with caution. We suggest using multiple i.i.d. replications of the supporting simulation model to confirm these properties and evaluate the confidence interval coverage. If these standard methods do not perform well for the supporting simulation models, then we can consider more sophisticated estimation methods, as in Alexopoulos et al. (2007), Tafazzoli et al. (2011), Tafazzoli and Wilson (2011) and references therein.

4.1. A Ratio Estimator

In both stationary and nonstationary settings, a CI (interval estimate) for $E[\overline{W}(t)]$ without observing the waiting times can be obtained using $\overline{W}_{L,\lambda}(t)$ if we can apply the following theorem, implementing the delta method; see Asmussen and Glynn (2007, §III.3 and Proposition §IV.4.1) for related results.

THEOREM 4 (ASYMPTOTICS FOR THE RATIO OF LOW-VARIABILITY POSITIVE NORMAL RANDOM VARIABLES). *If there is a sequence of systems indexed by n such that*

$$\sqrt{n}(\bar{L}^{(n)}(t) - L, \bar{\lambda}^{(n)}(t) - \lambda) \implies N(0, \Sigma) \text{ in } \mathbb{R}^2 \quad \text{as } n \to \infty, \tag{12}$$

where L and λ are positive real numbers and $N(0, \Sigma)$ is a mean-zero bivariate Gaussian random vector with variance vector $(\sigma_L^2, \sigma_\lambda^2)$ and covariance $\sigma_{L,\lambda}^2$, and $\overline{W}^{(n)}(t)$ satisfies

$$\overline{W}^{(n)}(t)/\overline{W}_{L,\lambda}^{(n)}(t) \implies 1 \quad \text{as } n \to \infty, \tag{13}$$

for $\overline{W}^{(n)}_{L,\lambda}(t) \equiv \bar{L}^{(n)}(t)/\bar{\lambda}^{(n)}(t)$, *then*

$$\sqrt{n}(\overline{W}^{(n)}(t) - (L/\lambda)) \Longrightarrow N(0, \sigma_W^2) \text{ in } \mathbb{R} \quad \text{as } n \to \infty \tag{14}$$

for

$$\sigma_W^2 = \frac{1}{\lambda^2}\left(\sigma_L^2 - \frac{2L\sigma^2_{L,\lambda}}{\lambda} + \frac{L^2\sigma_\lambda^2}{\lambda^2}\right). \tag{15}$$

Proof. Apply a Taylor expansion with the function $f(x, y) \equiv x/y$, having first partial derivatives $f_x = 1/y$ and $f_y = -x/y^2$, to get

$$\frac{\bar{L}^{(n)}(t)}{\bar{\lambda}^{(n)}(t)} = \frac{L}{\lambda} + \frac{\bar{L}^{(n)}(t) - L}{\lambda} - \frac{L(\bar{\lambda}^{(n)}(t) - \lambda)}{\lambda^2} + o\big(\max\{|\bar{L}^{(n)}(t) - L|, |\bar{\lambda}^{(n)}(t) - \lambda|\}\big), \tag{16}$$

so that

$$\sqrt{n}\big(\overline{W}^{(n)}_{L,\lambda}(t) - (L/\lambda)\big) = \frac{\sqrt{n}(\bar{L}^{(n)}(t) - L)}{\lambda} - \frac{\sqrt{n}L(\bar{\lambda}^{(n)}(t) - \lambda)}{\lambda^2} + o(1) \quad \text{as } n \to \infty, \tag{17}$$

from which (14) follows, given (12) and (13). □

We can apply the theorem if our system can be regarded as system n for n sufficiently large that we can replace the limits with approximate equality. The approximate confidence interval estimate for $E[\overline{W}^{(n)}(t)]$ would then be $[\overline{W}^{(n)}_{L,\lambda}(t) - 1.96\sigma_W/\sqrt{n}, \overline{W}^{(n)}_{L,\lambda}(t) + 1.96\sigma_W/\sqrt{n}]$, where σ_W is the square root of the variance σ_W^2 in (15). Since the variance σ_W^2 in (15) is typically unknown, we must estimate it. That can be done by inserting estimates for all the components of (15). Assuming that the estimates converge as $n \to \infty$, we still have asymptotic normality with the estimated values of the variance σ_W^2.

The sequence of systems indexed by n satisfying condition (12) in Theorem 4 can arise in two natural ways: First, condition (12) is typically satisfied if the averages are collected from a single observation over successively longer time intervals in a stationary environment, i.e., if t is allowed to grow with n, with $t_n \to \infty$. Then, of course, $E[\overline{W}^{(n)}(t)] \to W$ as $n \to \infty$, and we are simply estimating W. Second, whether or not there is a stationary environment, condition (12) is satisfied if the averages indexed by n correspond to averages taken over n multiple independent samples for a fixed interval $[0, t]$. The second case is important for the common case of service systems with strongly time-varying arrival rates over each day, provided that multiple days can be regarded as i.i.d. samples.

Condition (13) in Theorem 4 is of course also satisfied if the averages are collected from a single observation over successively longer time intervals in a stationary environment. However, condition (13) may well *not* be satisfied, even approximately, if the averages indexed by n correspond to averages taken over n multiple independent samples for a fixed interval $[0, t]$, because the bias may be significant, and it does not go away with increasing n; see §5.

4.2. The Supporting Central Limit Theorem in a Stationary Setting

With one sample-path segment, we suggest applying the method of batch means. A partial basis for that is the central limit theorem (CLT) version of Little's Law in Glynn and Whitt (1986) and Whitt (2012). To apply it, we assume that the system is approximately stationary over the designated subinterval $[0, t]$. Hence we regard the finite averages in (1) as estimators of the unknown parameters L, λ, and W. The CLT states that, under very general regularity conditions,

$$\big(\hat{L}(t), \hat{\lambda}(t), \hat{W}(t), \hat{L}_{W,\lambda}(t), \hat{\lambda}_{L,W}(t), \hat{W}_{L,\lambda}(t)\big) \Longrightarrow (X_L, X_\lambda, X_W, X_L, X_\lambda, X_W) \text{ in } \mathbb{R}^6 \tag{18}$$

as $t \to \infty$, where

$$\begin{aligned}&\big(\hat{L}(t), \hat{\lambda}(t), \hat{W}(t)\big) \equiv \sqrt{t}\big(\bar{L}(t) - L, \bar{\lambda}(t) - \lambda, \overline{W}(t) - W\big),\\ &\big(\hat{L}_{W,\lambda}(t), \hat{\lambda}_{L,W}(t), \hat{W}_{L,\lambda}(t)\big) \equiv \sqrt{t}\big(\bar{L}_{W,\lambda}(t) - L, \bar{\lambda}_{L,W}(t) - \lambda, \overline{W}_{L,\lambda}(t) - W\big),\end{aligned} \tag{19}$$

with the averages given in (1) and (3), and the limiting random vector (X_L, X_λ, X_W) is an essentially two-dimensional mean-zero multivariate Gaussian random vector with $X_W = \lambda^{-1}(X_L - WX_\lambda))$, so that the variance and covariance terms are related by

$$\begin{aligned}\sigma_W^2 &\equiv \mathrm{Var}(X_W) = E[X_W^2]\\ &= \lambda^{-2}(\sigma_L^2 - 2W\sigma^2_{\lambda,L} + W^2\sigma_\lambda^2),\\ \sigma^2_{L,W} &\equiv \mathrm{Cov}(X_L, X_W) = E[X_L X_W]\\ &= \lambda^{-1}(\sigma_L^2 - W\sigma^2_{\lambda,L}),\\ \sigma^2_{W,\lambda} &\equiv \mathrm{Cov}(X_W, X_\lambda) = E[X_W X_\lambda]\\ &= \lambda^{-1}(\sigma^2_{\lambda,L} - W\sigma_\lambda^2).\end{aligned} \tag{20}$$

Note that σ_W^2 in (20) agrees with (15).

Under general regularity conditions (essentially, if $t^{1/2}T_W^{(r)}(t) \Rightarrow 0$ for $T_W^{(r)}(t)$ in (5)), a functional central limit theorem (FCLT) generalization of the joint CLT in (18) is valid if a FCLT is valid in $\mathbb{R}^2$ for any two of the first three components. For example, it suffices to start with (FCLT generalization of) the bivariate CLT

$$\sqrt{t}(\bar{L}(t) - L, \bar{\lambda}(t) - \lambda) \Longrightarrow (X_L, X_\lambda) \text{ in } \mathbb{R}^2 \quad \text{as } t \to \infty, \tag{21}$$

where the limit (X_L, X_λ) is a bivariate mean-zero Gaussian random vector with variances σ_λ^2, σ_L^2, and covariance $\sigma^2_{\lambda,L}$. Natural sufficient conditions are based on regenerative structure for the stochastic process $\{L(t)\colon t \geqslant 0\}$, as in Asmussen (2003, §VI.3) and Glynn and Whitt (1987).

We directly assume that the limit in (18) is valid, and discuss how to apply it. Note that condition (21) coincides with condition (12) in Theorem 4, but now the conclusion directly gives a CLT for $\overline{W}(t)$ as well as for $\overline{W}_{L,\lambda}(t)$.

The form of the limit in (18) implies that the alternative estimators $\bar{L}_{W,\lambda}(t)$, $\bar{\lambda}_{L,W}(t)$, and $\overline{W}_{L,\lambda}(t)$ in (3) not only converge to the same limits L, λ and W just as the natural estimators $\bar{L}(t)$, $\bar{\lambda}(t)$, and $\overline{W}(t)$ in (1) do, but also the corresponding CLT-scaled random variables are asymptotically equivalent as well, i.e., $\|(\hat{L}(t),\ \hat{\lambda}(t),\ \hat{W}(t)) - (\hat{L}_{W,\lambda}(t),\ \hat{\lambda}_{L,W}(t),\ \hat{W}_{L,\lambda}(t))\| \Rightarrow 0$ as $t \to \infty$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^3$.

In summary, the CLT version of $L = \lambda W$ implies that the asymptotic efficiency (halfwidth of confidence intervals for large sample sizes) is the same for the alternative estimators in (3) as it is for the natural estimators in (1) (in a stationary setting). However, if one of the parameters happens to be known in advance, one estimator can be more efficient than the other; see Glynn and Whitt (1989). For example, with simulation, the arrival rate is typically known in advance.

4.3. Estimating Confidence Intervals by the Method of Batch Means

Assuming that the conditions for the CLT in the previous section are satisfied, given the sample-path segments $\{(A(s), L(s)):\ 0 \leqslant s \leqslant t\}$ and $\{W_k:\ R(0)+1 \leqslant k \leqslant R(0)+A(t)\}$ over the time interval $[0,t]$ (or only two of these three segments), we can use m batches based on measurements over the m subintervals $[(k-1)t/m, kt/m]$, $1 \leqslant k \leqslant m$. To define the batch averages, let $R_k \equiv R(kt/m)$, the number of customers remaining in the system at time kt/m from among those that arrived previously. Let $\bar{A}_k(t,m)$, $\bar{L}_k(t,m)$, and $\overline{W}_k(t,m)$ denote the averages over the interval $[(k-1)t/m, kt/m]$, i.e.,

$$\begin{aligned}
&\bar{A}_k(t,m) \equiv (m/t)A_k(t,m),\\
&\bar{L}_k(t,m) \equiv (m/t)L_k(t,m),\\
&\overline{W}_k(t,m) \equiv (1/A_k(t,m))W_k(t,m),\\
&A_k(t,m) \equiv A(kt/m) - A((k-1)t/m),\\
&L_k(t,m) \equiv \textstyle\int_{(k-1)t/m}^{kt/m} L(s)\,ds, \quad W_k(t,m) \equiv \sum_{j=R_{k-1}+1}^{R_{k-1}+A_k(t,m)} W_j.
\end{aligned} \tag{22}$$

The FCLT version of the CLT in the previous section implies that, as $t \to \infty$, the vector of scaled batch means $\sqrt{t/m}(\bar{A}_k(t,m)-\lambda, \bar{L}_k(t,m)-L, \overline{W}_k(t,m)-W)$, $1 \leqslant k \leqslant m$, are asymptotically m i.i.d. mean-zero Gaussian random vectors with variances σ_λ^2, σ_L^2, and σ_W^2, and covariances $\sigma_{L,\lambda}^2$, $\sigma_{\lambda,W}^2$, and $\sigma_{L,W}^2$. By Theorem 4, as $t \to \infty$, the associated scaled vector $\sqrt{t/m}(\overline{W}_{L,\lambda,k}(t,m)-W)$, $1 \leqslant k \leqslant m$, are asymptotically m i.i.d. mean-zero random variables with variance σ_W^2 in (15). Hence, as $t \to \infty$, also

$$\frac{\sum_{k=1}^m (\overline{W}_{L,\lambda,k}(t,m) - \overline{W}_{L,\lambda}^{(m)}(t))}{\sqrt{S_{(m)}^2(t)/m}} \implies t_{m-1}, \tag{23}$$

where t_{m-1} is a random variable with the Student t distribution with $m-1$ degrees of freedom,

$$\begin{aligned}
&\overline{W}_{L,\lambda}^{(m)}(t) \equiv \frac{1}{m}\sum_{k=1}^m \overline{W}_{L,\lambda,k}(t,m) \quad \text{and}\\
&S_{(m)}^2(t) \equiv \frac{1}{m-1}\sum_{k=1}^m (\overline{W}_{L,\lambda,k}(t,m) - \overline{W}_{L,\lambda}^{(m)}(t))^2.
\end{aligned} \tag{24}$$

Thus,

$$\left[\overline{W}_{L,\lambda}^{(m)}(t) - \frac{t_{0.025,\,m-1}S_{(m)}(t)}{\sqrt{m}},\quad \overline{W}_{L,\lambda}^{(m)}(t) + \frac{t_{0.025,\,m-1}S_{(m)}(t)}{\sqrt{m}}\right]$$

is an approximate 95% confidence interval for W based on the t distribution and the average $\overline{W}_{L,\lambda}^{(m)}(t)$ of batch means. Of course, the same procedure applies to other averages of batch means as well.

It remains to choose the number of batches, m. Since we obtain larger batch sizes, and thus more nearly approximate the asymptotic condition $t \to \infty$, if we make m small, we advocate keeping it relatively small, e.g., $m = 5$. Nevertheless, in our examples we consider a range of m values.

5. Estimating and Reducing the Bias

We now discuss ways to estimate and reduce the bias in the estimator $\overline{W}_{L,\lambda}(t)$ in (3) as an estimator for $E[\overline{W}(t)]$ for $\overline{W}(t)$ in (1). In doing so, we are primarily concerned with nonstationary settings. In stationary settings, $\overline{W}(t)$ in (1) is typically a biased estimator of W, whereas $\overline{W}_{L,\lambda}(t)$ is typically a biased estimator of both W and $E[\overline{W}(t)]$, but these biases are less likely to be serious, e.g., see §5.4.

An important conclusion from our analysis is that the bias depends on the underlying model. We demonstrate by considering two idealized paradigms: the infinite-server and single-server paradigms. We emphasize the infinite-server paradigm, which often is appropriate for call centers. In §5.4, we show that the bias in $\overline{W}(t)$ for estimating W tends to be negligible in the infinite-server paradigm.

5.1. Bias in $\overline{W}_{L,\lambda}(t)$ as an Estimator of the Expected Average Wait $E[\overline{W}(t)]$

Since the bias in $\overline{W}_{L,\lambda}(t)$ as an estimator for $E[\overline{W}(t)]$ is $E[\Delta_W(t)]$ for $\Delta_W(t) \equiv \overline{W}_{L,\lambda}(t) - \overline{W}(t)$ in (7), we can apply Theorem 2 to obtain an exact expression for the bias $E[\Delta_W(t)]$. We also give the conditional bias $E[\Delta_W(t) \mid \mathcal{O}_t]$ given the observed data over the interval $[0,t]$, which we assume is $\mathcal{O}(t) \equiv (t, \bar{L}(t), \bar{\lambda}(t), R(0), L(t))$, from which we can also deduce $A(t)$. We use the conditional bias to create a refined estimator given the observed data.

Corollary 1 (Exact Bias and Conditional Bias). *The bias in $\overline{W}_{L,\lambda}(t)$ in (3) as an estimator for $E[\overline{W}(t)]$ for $\overline{W}(t)$ in (1) is $E[\Delta_W(t)] = E[E[\Delta_W(t) \mid \mathcal{O}(t)]]$, where $\Delta_W(t)$ is given in (7), the vector of observed data is $\mathcal{O}(t) \equiv (t, \bar{L}(t), \bar{\lambda}(t), R(0), L(t))$ and the conditional bias is*

$$E[\Delta_W(t) \mid \mathcal{O}_t] = \frac{\sum_{k=1}^{R(0)} E[W_k^{r,0} \mid \mathcal{O}_t] - \sum_{k=1}^{L(t)} E[W_k^{r,t} \mid \mathcal{O}_t]}{A(t)}. \tag{25}$$

Proof. Apply Theorem 2 using (5). □

5.2. Two Approximations

The bias in Corollary 1 is not easy to analyze. Given that $(R(0), L(t), A(t))$ is observed, it remains to estimate the conditional residual waiting times $E[W_k^{r,0} \mid \mathcal{O}_t]$, $1 \leqslant k \leqslant R(0)$, and $E[W_k^{r,t} \mid \mathcal{O}_t]$, $1 \leqslant k \leqslant L(t)$. The conditional expectations $E[W_k^{r,0} \mid \mathcal{O}_t]$ are complicated, because we are conditioning on events in the future after the observation time 0. Thus, we develop two approximations and then show that they apply to the infinite-server paradigm.

5.2.1. Simplification from the Bias Approximation Assumption. As t increases, we expect the "initial edge effect" $\{R(0), W_k^{r,0}; 1 \leqslant k \leqslant R(0)\}$ to be approximately independent of the "terminal edge effect" $\{L(t), W_k^{r,t}; 1 \leqslant k \leqslant L(t)\}$ and the total number of arrivals $A(t)$. With that in mind, we use the following approximation, which primarily means that we are assuming that t is sufficiently large.

Bias Approximation Assumption (*BAA*). For $\mathcal{O}(t) \equiv (t, \bar{L}(t), \bar{\lambda}(t), R(0), L(t))$, $t \geqslant 0$,

$$E[W_k^{r,0} \mid \mathcal{O}_t] \approx E[W_k^{r,0} \mid R(0)], \quad 0 \leqslant k \leqslant R(0), \quad \text{and}$$
$$E[W_k^{r,t} \mid \mathcal{O}_t] \approx E[W_k^{r,t} \mid L(t)], \quad 0 \leqslant k \leqslant L(t).$$

Invoking the BAA, we obtain the following approximation directly from Corollary 1:

$$E[\Delta_W(t) \mid \mathcal{O}_t] \approx \frac{\sum_{k=1}^{R(0)} E[W_k^{r,0} \mid R(0)] - \sum_{k=1}^{L(t)} E[W_k^{r,t} \mid L(t)]}{A(t)}. \tag{26}$$

We think that BAA is reasonable if t is sufficiently large. That is easy to see for stationary models, because then as $t \to \infty$ (i) $\bar{L}(t) \to L$ and $\bar{\lambda}(t) \to \lambda$ and (ii) under regularity conditions (e.g., regenerative structure), $\{R(0), W_k^{r,0}; 1 \leqslant k \leqslant R(0)\}$ will be asymptotically independent of $\{L(t), W_k^{r,t}; 1 \leqslant k \leqslant L(t)\}$.

5.2.2. Using $\overline{W}_{L,\lambda}(t)$ to Estimate the Residual Waiting Times. We can obtain an applicable estimate of the conditional bias $E[\Delta_W(t) \mid \mathcal{O}_t]$ in (25) if we estimate all the remaining conditional waiting times by the observed $\overline{W}_{L,\lambda}(t)$. In doing so, we are ignoring the inspection paradox (since these are remainders of waiting times in progress), the model structure and the available information $\mathcal{O}(t)$. This step is likely to be justified approximately if the distribution of the waiting times is nearly exponential.

That step yields the approximation

$$E[\Delta_W(t) \mid \mathcal{O}_t] \approx \frac{(R(0) - L(t))\overline{W}_{L,\lambda}(t)}{A(t)} \quad \text{for}$$
$$\mathcal{O}(t) \equiv (R(0), L(t), \bar{L}(t), \bar{\lambda}(t)). \tag{27}$$

We can apply approximation (27) to obtain the new candidate *refined estimator* of $E[\overline{W}(t)]$, exploiting the observed vector $(R(0), L(t), A(t))$:

$$\overline{W}_{L,\lambda,r}(t) \equiv \overline{W}_{L,\lambda}(t) - E[\Delta_W(t) \mid \mathcal{O}_t] \approx \overline{W}_{L,\lambda}(t)\left(1 - \frac{R(0) - L(t)}{A(t)}\right). \tag{28}$$

(The refined estimator $\overline{W}_{L,\lambda,r}(t)$ in (28) is a candidate refinement of the indirect estimator $\overline{W}_{L,\lambda}(t)$ (3).) The associated approximate relative conditional bias is thus

$$E[\Delta_W^{\text{rel}}(t) \mid \mathcal{O}(t)] \equiv \frac{E[\Delta_W(t) \mid \mathcal{O}_t]}{E[\overline{W}(t)]} \approx \frac{E[\Delta_W(t) \mid \mathcal{O}_t]}{\overline{W}_{L,\lambda}(t)} \approx \frac{R(0) - L(t)}{A(t)}. \tag{29}$$

In the next section we show that the analysis in (27)–(29) can be supported theoretically in the infinite-server paradigm when the waiting times are exponential, so we propose the refined estimator in (28) as a candidate estimator for many-server systems. However, the crude analysis above is not justified universally; e.g., it is not good for the single-server models, as we show in §5.5.

5.3. The Infinite-Server Paradigm

If, in addition to BAA, we consider the $G_t/M/\infty$ IS model with exponential service times having mean $E[S]$, then (26) becomes

$$E[\Delta_W(t) \mid \mathcal{O}_t] \approx (R(0) - L(t))E[S]/A(t). \tag{30}$$

Since the waiting times coincide with the service times in the IS model, it is natural to use the observed $\overline{W}_{L,\lambda}(t)$ as an initial estimate of $E[S]$. If we use $\overline{W}_{L,\lambda}(t)$ as an estimate of $E[S]$ in (30), then the formula in (30) reduces to the bias approximation in (27). Thus, under these approximations, the refined estimator (28) becomes unbiased. Hence, we propose the refined estimator in (28) for light to moderately loaded many-server systems with service time-distributions not too far from exponential.

To better understand the consequence of nonexponential service times in the infinite-server paradigm, we now consider the $M_t/GI/\infty$ IS model with nonexponential service times. We assume that it starts empty at some time in the past (possibly in the infinite past) having bounded time-varying arrival rate $\lambda(t)$, i.i.d. service times, independent of the arrival process, with generic service-time S having cdf $G(x) \equiv P(S \leqslant x)$ with $E[S^2] < \infty$ and thus the finite-squared coefficient of variation (SCV) $c_S^2 \equiv \text{Var}(S)/E[S]^2$. Let $G^c(x) \equiv 1 - G(x)$ be the complementary cdf. Let S_e be an associated random variable with the associated stationary excess or residual lifetime distribution,

$$P(S_e \leqslant x) \equiv \frac{1}{E[S]} \int_0^x G^c(u)\, du \quad \text{and}$$
$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}. \tag{31}$$

For this IS model, we can characterize the conditional expected value of the remaining work $T_W^{(r)}(t)$ in (5) and (7) given $L(t)$, but it requires the full waiting-time cdf G.

THEOREM 5 (TOTAL REMAINING WORK FOR THE $M_t/GI/\infty$ INFINITE-SERVER MODEL). *For the $M_t/GI/\infty$ model above,*

$$E[T_W^{(r)}(t) \mid L(t)] = \frac{L(t)\int_0^\infty \lambda(t-u)E[S-u; S>u]\,du}{E[L(t)]}, \tag{32}$$

for $T_W^{(r)}(t)$ in (5), where $E[S-u; S>u] = E[S-u \mid S>u]P(S>u)$, $E[S-u \mid S>u] = \int_0^\infty (G^c(x-u)/G^c(u))\,dx$, and

$$E[L(t)] = \int_0^\infty \lambda(t-u)G^c(u)\,du = E[\lambda(t-S_e)]E[S], \quad t \geqslant 0. \tag{33}$$

PROOF. Conditional on $L(t)=n$, the n customers remaining in service have i.i.d. service times distributed as S_t with

$$P(S_t > x) = \frac{\int_0^\infty \lambda(t-u)P(S>x+u)\,du}{E[L(t)]}, \tag{34}$$

for $E[L(t)]$ given in (33), by Goldberg and Whitt (2008, Theorem 2.1), which draws on Eick et al. (1993a). There the system starts empty at time 0, but the result extends to the present setting, given that we have assumed that the arrival rate function is bounded and $E[S^2] < \infty$. The second expression in (33) is given in Eick et al. (1993a, Theorem 1). □

If we now invoke the BAA for the $M_t/GI/\infty$ model, then we obtain the approximation

$$E[\Delta_W(t) \mid \mathcal{O}_t] \approx \frac{E[T_W^{(r)}(0) \mid L(0)] - E[T_W^{(r)}(t) \mid L(t)]}{A(t)}, \tag{35}$$

where (32) can be used to compute both terms in the numerator.

In practice, we presumably would not know the full service-time cdf, so that the approximation in (35) based on Theorem 5 would not appear to be very useful, but we now show that it provides strong support for the refined estimator in (28) if the service time is not too far from exponential. For that purpose, we observe that the complicated formula above simplifies in special cases. First, for $M_t/M/\infty$, formula (32) reduces to

$$E[T_W^{(r)}(t) \mid L(t)] = L(t)E[S],$$

taking us back to (27). Second, for the stationary $M/GI/\infty$ model starting empty in the infinite past, S_t in (34) is distributed as S_e in (31), so that formula (32) reduces to

$$E[T_W^{(r)}(t) \mid L(t)] = L(t)E[S_e] = L(t)E[S](c_s^2+1)/2$$

and (35) reduces to

$$E[\Delta_W(t) \mid \mathcal{O}_t] = (R(0) - L(t))E[S](c_s^2+1)/2A(t),$$

depending only on the first two moments of the distribution.

This result for the stationary $M/GI/\infty$ model applies to the nonstationary $M_t/GI/\infty$ system if the arrival rate is nearly constant just prior to the two times 0 and t, where we would be applying Theorem 5. Thus, we conclude that this section provides strong support for the refined estimator $\overline{W}_{L,\lambda,r}(t)$ in (28) in the common case where (i) the arrival rate changes relatively slowly compared to the mean service time and (ii) the service-time SCV c_s^2 is not too far from 1, as is often the case in call centers, e.g., here (where $c_s^2 = 1.017$) and in Brown et al. (2005). We could obtain a further refinement if we could estimate the SCV c_s^2.

5.4. Bias of $\overline{W}(t)$ in the Infinite-Server Paradigm

We now observe that the bias of $\overline{W}(t)$ as an estimator of W should usually not be a major factor in the infinite-server paradigm. We do so by showing that the bias is quantifiably small for an IS model. We use the $G_t/GI/\infty$ model with general, possibly nonstationary, arrival counting process A. The key assumption is that the waiting times, which coincide with the service times, are i.i.d. with mean W and independent of the arrival process. Using that independence, we can write

$$E[\overline{W}(t) \mid A(t)>0] = E[E[\overline{W}(t) \mid A(t)] \mid A(t)>0] = E[W \mid A(t)>0] = W. \tag{36}$$

Given that we have defined $\overline{W}(t) \equiv 0$ when $A(t)=0$, we have the following result.

THEOREM 6 (CONDITIONAL BIAS OF THE AVERAGE WAITING TIME IN THE $G_t/GI/\infty$ MODEL). *For the $G_t/GI/\infty$ infinite-server model, having i.i.d. service times with mean W, that are independent of a general arrival process,*

$$E[\overline{W}(t)] = WP(A(t)>0). \tag{37}$$

For a stationary Poisson arrival process with rate λ, $W - E[\overline{W}(t)] = We^{-\lambda t}$, $t \geqslant 0$.

5.5. The Single-Server Paradigm

To show that the refined estimator $\overline{W}_{L,\lambda,r}(t)$ in (28) is not always good and that the bias can be analyzed exactly in some cases and can be significant, we now consider a single-server model. Let $L(t)$ be the number of customers waiting in queue in a single-server $G_t/GI/1$ queueing model with unlimited waiting space and the first-come first-served service discipline, with a general arrival process possibly having a time-varying arrival-rate function $\lambda(t)$ and service times S_i that are independent and identically distributed (i.i.d.) and independent of the arrival process, each distributed as a random variable S having cdf $G(x)$. In addition to the model structure, we assume that we know the mean $E[S]$, which in practice may be based on a sample mean estimate.

We now assume that $T_W^{(r)}(0)$ in (5) is observable, which is reasonable because customers depart in order of arrival in the single-server model. It is also necessary for all these customers to have departed by time t, which is reasonable if t is not too small. Let $S^{(r)}(t)$ be the residual service time

of the customer in service at time t, if any, In this setting, the total remaining waiting time of all customers in the system at time t is given by

$$T_W^{(r)}(t) \equiv \sum_{k=1}^{L(t)} W_k^{r,t} = L(t)S^{(r)}(t) + \sum_{k=1}^{L(t)-1} (L(t)-k)S_{k+1}, \quad (38)$$

where S_k, $k \geqslant 2$, are i.i.d. and independent of $L(t)$ and $S^{(r)}(t)$, but in general $L(t)$ and $S^{(r)}(t)$ are dependent. Further simplification occurs if S is exponential.

THEOREM 7 (BIAS REDUCTION FOR THE $G_t/M/1/\infty$ MODEL). *For the $G_t/M/1/\infty$ model,*

$$E[T_W^{(r)}(t) \mid L(t), E[S]] = L(t)(L(t)+1)E[S]/2, \quad (39)$$

so that, if $T_W^{(r)}(0)$ is fully observable in $[0,t]$, then

$$E[\Delta_W(t) \mid \mathcal{O}_t] = \frac{T_W^{(r)}(0) - L(t)(L(t)+1)E[S]/2}{A(t)} \quad \textit{for}$$

$$\mathcal{O}(t) \equiv (L(t), \bar{L}(t), \bar{\lambda}(t), T_W^{(r)}(0), E[S]). \quad (40)$$

PROOF. Formula (40) follows directly from (39), which in turn follows from (38) given that $S^{(r)}(t)$ has the same exponential distribution as S_1 and $1 + \cdots + (n-1) = n(n-1)/2$. □

We apply Theorem 7 to obtain the single-server refined estimator

$$\overline{W}_{L,\lambda,r,1}(t) \equiv \overline{W}_{L,\lambda}(t) - \frac{T_W^{(r)}(0) - L(t)(L(t)+1)E[S]/2}{A(t)}. \quad (41)$$

Even if we do not know the mean $E[S]$, formulas (39)–(41) provide important insight, showing that $E[T_W^r(t) \mid L(t), E[S]]$ is approximately proportional to $L(t)^2$ instead of $L(t)$ as in (27) and §5.3. We next show that the bias in (40) can be significant by considering a transient $M/M/1$ example.

A Simulation Example: The $M/M/1$ Queue Starting Empty. To illustrate the bias for single-server models discussed in §5.5, we report results from a simulation experiment for the $M/M/1$ queue with mean service time $1/\mu = 1$ starting empty over the interval $[0, 10]$ for three values of the constant arrival rate λ: 0.7, 1.0 and 2.0. The respective 95% confidence intervals (CIs) for the exact value of $E[\overline{W}(t)]$ estimated by the sample average of 1,000 replications of $\overline{W}(t)$ were 1.88 ± 0.08, 2.70 ± 0.12, and 6.36 ± 0.19; the sample means are regarded as the exact values. (In an application of Little's Law, these direct estimates would not be available.) To see that the refined estimator $\overline{W}_{L,\lambda,r,1}(t)$ in (41) has essentially no bias at all, without expense of wider CIs, the corresponding CIs for it based on the same 1,000 replications were 1.90 ± 0.08, 2.68 ± 0.11 and 6.38 ± 0.18. In contrast, the unrefined $\overline{W}_{L,\lambda}(t)$ in (3) produced the corresponding tighter erroneous CIs 1.47 ± 0.06, 1.82 ± 0.06 and 2.85 ± 0.06. From the analysis above, we should not expect that the $G_t/M/\infty$ refined estimator (28) should perform well here. That is confirmed by the corresponding CIs 1.83 ± 0.08, 2.46 ± 0.10 and 4.46 ± 0.11. That is pretty good for $\lambda = 0.7$, but it misses badly for $\lambda = 2.0$.

6. Confidence Intervals for the Refined Estimator

We now see how the two statistical techniques in §§4 and 5 can be combined. We estimate confidence intervals for the refined estimators in (28) and (41), as well as the other estimators in (1) and (3).

6.1. Confidence Intervals for the Mean Wait in the Transient $M/M/1$ Queue

We now give an example in which *both* bias reduction and estimating confidence intervals contribute significantly to our understanding. To see large bias, we return to the example of the transient $M/M/1$ queue in §5.5. We now show how the sample average approach can be applied to estimate confidence intervals for the refined estimator in (41) that eliminates the bias. We now consider 10 i.i.d. samples of the same $M/M/1$ model over the interval $[0, 10]$, starting empty at time 0. We study the CI coverage by performing 1,000 replications of the entire experiment.

Table 4 shows that the unrefined estimator $\overline{W}_{L,\lambda}(t)$ in (3) does a very poor job in estimating the mean wait because of the bias, but the performance of the refined estimator $\overline{W}_{L,\lambda,r}(t)$ in (28) and the direct estimator $\overline{W}(t)$ is not too bad. It is known that residual skewness of the estimates can degrade the performance of confidence intervals, but we find that our estimates are not extreme examples of nonnormality and skewness; see §4 of the e-companion for details. In an effort to obtain a better estimate of confidence intervals, one can consider using the appropriate confidence interval inflation factor. We estimate it to be about 1.55, 1.45, and 1.05 for $\lambda = 0.7$, 1.0, and 2.0, respectively (details in §4 of the e-companion). For more discussion on skewness-adjusted CI, see Johnson (1978) and Willink (2005); in the context of batch means and their residual skewness and correlations, see Alexopoulos and Goldsman (2004), Tafazzoli et al. (2011), Tafazzoli and Wilson (2011), and references therein.

6.2. Evaluating the Refined Estimator with the Call Center Data

Given that the call center should approximately fit the infinite-server paradigm and that the waiting-time distribution is approximately exponential, we can apply Equation (29) to see that the bias should be relatively small in the call center example. We now use data from the 18 weekdays in May 2001 for the call center example in §3 to confirm that observation and show that the refined estimator in (28) is effective in reducing the bias.

Since we observe strong day-to-day variation in the average waiting times, we do not try to estimate the overall

Table 4. Confidence intervals for the mean wait in the transient $M/M/1$ queue for $\lambda = 0.7, 1.0,$ and 2.0.

λ	$\bar{L}(t)$	$\bar{\lambda}(t)$	$\overline{W}(t)$	Cov. (%)	$\overline{W}_{L,\lambda}(t)$	Cov. (%)	$\overline{W}_{L,\lambda,r}(t)$	Cov. (%)
0.7	1.10 ± 0.57	0.70 ± 0.17	1.89 ± 0.85	90.3	1.46 ± 0.57	58.3	1.88 ± 0.82	89.9
1.0	1.91 ± 0.88	1.00 ± 0.21	2.63 ± 1.16	90.5	1.80 ± 0.63	31.5	2.62 ± 1.11	92.2
2.0	5.82 ± 1.74	2.00 ± 0.29	6.36 ± 2.07	91.3	2.83 ± 0.63	0.0	6.38 ± 1.95	93.1

Notes. Results are based on 1,000 replications of 10 i.i.d. samples of the same $M/M/1$ model over the interval $[0, 10]$, starting empty at time 0. True mean wait values are estimated using 100,000 simulation runs and assumed to be 1.8913, 2.6354, and 6.3786 for $\lambda = 0.7$, 1.0, and 2.0, respectively.

Table 5. Comparison of the refined estimator $\overline{W}_{L,\lambda,r}(t)$ in (28) to the unrefined estimator $\overline{W}_{L,\lambda}(t)$ in (3): Average over the day of the average absolute errors (AAE) and average squared errors (ASE) for each time interval over 18 weekdays in the call center example.

Subinterval length	Intervals averaged over	Unrefined in (3) $\overline{W}_{L,\lambda}(t)$	AAE	ASE	Refined in (28) $\overline{W}_{L,\lambda,r}(t)$	AAE	ASE
Hours	[6, 10]	3.32	0.241	0.117	3.54	0.082	0.018
	[10, 16]	3.61	0.076	0.010	3.60	0.058	0.006
	[16, 23]	4.46	0.271	0.160	4.28	0.153	0.057
	All	3.89	0.195	0.097	3.86	0.103	0.030
Half hours	[6, 10]	3.27	0.303	0.198	3.49	0.169	0.068
	[10, 16]	3.62	0.161	0.052	3.60	0.110	0.020
	[16, 23]	4.55	0.533	0.673	4.25	0.340	0.322
	All	3.92	0.347	0.342	3.84	0.219	0.156

mean over all days, but aim to estimate the mean of specified intervals on each day (for sample averages over all days and their associated confidence interval, see §6.3). In particular, we compute the average over the 18 days of the absolute errors $|\overline{W}_{L,\lambda}(t) - \overline{W}(t)|$ (AAE) and associated average squared errors (ASE) for each of the 17 hours and 34 half hours of the day. We choose hours and half hours, because they represent typical staffing intervals in call centers; see Green et al. (2007). Table 5 highlights the results; AAE and ASE of each subinterval (hours and half hours) are again averaged over the intervals [6, 10], [10, 16], [16, 23], and all day. More details appear in Kim and Whitt (2012, Tables 15 and 16).

Table 5 shows that the refined estimator reduces the AAE from 0.195 (about 5.0% of the overall average wait, 3.89) to 0.103 (2.6%) for hours over all hours, while the refined estimator reduces the AAE from 0.347 (8.9%) to 0.219 (5.6%) for half hours over all half hours. In both cases, there is more bias and more bias reduction at the ends of the day when the system is nonstationary. In addition, we note that the unrefined estimator underestimates $\overline{W}(t)$ during [6, 10] when the arrival rate is increasing, and that it overestimates $\overline{W}(t)$ during [16, 23] when the arrival rate is decreasing, as expected.

6.3. Sample Averages Over Separate Days

For many service systems, whether stationary or not, we may be able to estimate CIs for $E[\overline{W}(t)]$ in (1) without observing the waiting times via $E[\overline{W}_{L,\lambda}(t)]$ in (3) using sample averages over multiple days, regarding those days as approximately i.i.d. We assume that the time average operation makes the vector $(\bar{L}(t), \bar{\lambda}(t))$ approximately Gaussian for each day. Thus, by Theorem 4, the associated random variable $\overline{W}_{L,\lambda}(t)$ should be approximately Gaussian as well with (unknown) variance given in (15). We also assume that any refinement $\overline{W}_{L,\lambda,r}(t)$ is approximately Gaussian as well.

Based on n days regarded as i.i.d., we can construct CI in the usual way. Let X_i denote the time average $\overline{W}_{L,\lambda}(t)$ or (preferably) its refinement $\overline{W}_{L,\lambda,r}(t)$ based on the bias analysis described in §5 for day i. Let the sample mean and variance be

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{42}$$

Then $(\bar{X}_n - E[\overline{W}(t)])/\sqrt{S_n^2/n}$ should be approximately distributed as t_{n-1}, Student t with $n-1$ degrees of freedom. Then $\bar{X}_n \pm t_{\alpha/2,n-1} S_n/\sqrt{n}$ is a $1-\alpha$ CI for $E[\overline{W}(t)]$.

To assess how well indirect estimators perform in estimating $E[\overline{W}(t)]$ over separate days and in different settings, we again consider our call center data and divide each day into three intervals, [6, 10], [10, 16], and [16, 23] so that the arrival rate is increasing in [6, 10], approximately stationary in [10, 16], and decreasing in [16, 23]. The performance of two indirect estimators, the refined estimator $\overline{W}_{L,\lambda,r}(t)$ in (28) and the unrefined estimator $\overline{W}_{L,\lambda}(t)$ in (3), as well as that of the direct estimator, is illustrated in Table 6. (Additional estimation results appear in Kim and Whitt 2012, Tables 17–19.) We see that the refined estimator $\overline{W}_{L,\lambda,r}(t)$ behaves very similarly to the direct estimator in all cases. The unrefined estimator performs well in the stationary region [10, 16], but shows the impact of bias in nonstationary regions, [6, 10] and [16, 23], as expected.

Table 6. Estimating $E[\overline{W}(t)]$ and its associated 95% confidence interval over 18 weekdays in the call center example: Comparison of the refined estimator $\overline{W}_{L,\lambda,r}(t)$ in (28) to the unrefined estimator $\overline{W}_{L,\lambda}(t)$ in (3).

Intervals	Direct estimator $\overline{W}(t)$	Unrefined in (3) $\overline{W}_{L,\lambda}(t)$	Refined in (28) $\overline{W}_{L,\lambda,r}(t)$
[6, 10]	3.47 ± 0.22	3.35 ± 0.23	3.47 ± 0.23
[10, 16]	3.60 ± 0.11	3.61 ± 0.11	3.60 ± 0.11
[16, 23]	4.24 ± 0.26	4.35 ± 0.26	4.22 ± 0.25

7. Conclusions

Little's Law is an important theoretical cornerstone of operations research, but it does not apply directly to applications involving measurements over finite-time intervals. As reviewed in §2.3, it is possible to modify the definitions so that the relation $\bar{L}(t) = \bar{\lambda}(t)\overline{W}(t)$ always holds for finite averages, but we advocate not doing so. Instead, we advocate taking a statistical approach, estimating confidence intervals (§4) and considering modified estimators that reduce bias (§5), which exploit the extended finite-time Little's Law in Theorem 2. We have illustrated the statistical approach by applying it to the call center example in §3. We have focused on the problem of estimating the unknown mean values W and $E[\overline{W}(t)]$, using $\overline{W}_{L,\lambda}(t) \equiv \bar{L}(t)/\bar{\lambda}(t)$ when the waiting times cannot be directly observed.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2013.1193.

Acknowledgments

The authors thank Avishai Mandelbaum, Galit Yom-Tov, Ella Nadjharov, and the Center for Service Enterprise Engineering (SEE) at the Technion for access to the SEE call center data and advice about its use. They thank the Samsung Foundation and the National Science Foundation for support [NSF Grant CMMI 1066372].

References

Alexopoulos C, Goldsman D (2004) To batch or not to batch. *ACM Trans. Modeling Comput. Simulation (TOMACS)* 14:76–114.
Alexopoulos C, Argon NT, Goldsman D, Steiger NM, Tafazzoli A, Wilson JR (2007) Efficient computation of overlapping variance estimators for simulation. *INFORMS J. Comput.* 19:314–327.
Asmussen S (2003) *Applied Probability and Queues*, 2nd ed. (Springer, New York).
Asmussen S, Glynn PW (2007) *Stochastic Simulation: Algorithms and Analysis* (Springer, New York).
Baccelli F, Bremaud P (2003) *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences*, 2nd ed. (Springer, New York).
Brockwell PJ, Davis RA (1991) *Time Series Theory and Methods*, 2nd ed. (Springer, New York).
Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100:36–50.
Buzen JP (1976) Fundamental operational laws of computer system performance. *Acta Informatica* 7:167–182.
Denning PJ, Buzen JP (1978) The operational analysis of queueing network models. *Comput. Surveys* 10:225–261.
Eick SG, Massey WA, Whitt W (1993a) The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41:731–742.
Eick SG, Massey WA, Whitt W (1993b) $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* 39:241–252.
El-Taha M, Stidham S Jr (1999) *Sample-Path Analysis of Queueing Systems* (Kluwer, Boston).
Glynn PW, Whitt W (1986) A central-limit-theorem version of $L = \lambda W$. *Queueing Systems* 1:191–215.
Glynn PW, Whitt W (1987) Sufficient conditions for functional-limit-theorem versions of $L = \lambda W$. *Queueing Systems* 1:279–287.
Glynn PW, Whitt W (1989) Indirect estimation via $L = \lambda W$. *Oper. Res.* 37:82–103.
Goldberg DA, Whitt W (2008) The last departure time from an $M_t/G/\infty$ queue with a terminating arrival process. *Queueing Systems* 58:77–104.
Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16:13–39.
Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42:1383–1394.
Jewell WS (1967) A simple proof of $L = \lambda W$. *Oper. Res.* 15:1109–1116.
Johnson NJ (1978) Modified t tests and confidence intervals for asymmetrical populations. *J. Amer. Statist. Assoc.* 73:536–544.
Kim S-H, Whitt W (2012) Statistical analysis with Little's Law, supplementary material. Technical report, Columbia University, New York. http://www.columbia.edu/~ww2040/allpapers.htm.
Kim S-H, Whitt W (2013) Estimating waiting times with the time-varying Little's Law. *Probab. Engrg. Inform. Sci.* Forthcoming.
Little JDC (1961) A proof of the queueing formula: $L = \lambda W$. *Oper. Res.* 9:383–387.
Little JDC (2011) Little's Law as viewed on its 50th anniversary. *Oper. Res.* 59:536–539.
Little JDC, Graves SC (2008) Little's Law. Chhajed D, Lowe TJ, eds. *Building Intuition: Insights from Basic Operations Management Models and Principles*, Chap. 5 (Springer, New York), 81–100.
Lovejoy WS, Desmond JS (2011) Little's Law flow analysis of observation unit impact and sizing. *Acad. Emergency Medicine* 18:183–189.
Mandelbaum A (2011) Little's Law over a finite horizon. Teaching notes on Little's Law in a course on service engineering, October, 17.1–17.6. Accessed August 3, http://iew3.technion.ac.il/serveng/Lectures/lectures.html.
Mandelbaum A (2012) Service Engineering of Stochastic Networks Web Page. Accessed July 2013, http://iew3.technion.ac.il/serveng/.
Serfozo R (1999) *Introduction to Stochastic Networks* (Springer, New York).
Sigman K (1995) *Stationary Marked Point Processes, An Intuitive Approach* (Chapman and Hall, New York).
Stidham S Jr (1974) A last word on $L = \lambda W$. *Oper. Res.* 22:417–421.
Tafazzoli A, Wilson JR (2011) Skart: A skewness- and autoregression-adjusted batch-means procedure for simulation analysis *IIE Trans.* 43:110–128.
Tafazzoli A, Steiger NM, Wilson JR (2011) N-Skart: A nonsequential skewness- and autoregression-adjusted batch-means procedure for simulation analysis *IEEE Trans. Automatic Control* 56:254–264.
Whitt W (2012) Extending the FCLT version of $L = \lambda W$. *Oper. Res. Lett.* 40:230–234.
Willink R (2005) A confidence interval and test for the mean of an asymmetric distribution. *Comm. Statist. Theory Methods* 34:753–766.

Song-Hee Kim is a doctoral student in the Department of Industrial Engineering and Operations Research at Columbia University. Her primary research focus is on operations management in service systems with emphasis on problems related to healthcare delivery, using empirical/statistical analysis, simulation, stochastic modeling, and queueing theory.

Ward Whitt is a professor in the Department of Industrial Engineering and Operations Research at Columbia University. He joined the faculty in 2002 after spending 25 years in research at AT&T. He received his Ph.D. from Cornell University in 1969. He is a longtime member of INFORMS and its Applied Probability Society. His recent research has focused on stochastic models of service systems, using both queueing theory and simulation.