Data-driven Decisions in Service Systems

Song-Hee (Hailey) Kim

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

Abstract

Data-driven Decisions in Service Systems

Song-Hee (Hailey) Kim

This thesis makes contributions to help provide data-driven (or evidence-based) decision support to service systems, especially hospitals. Three selected topics are presented.

First, we discuss how Little's Law, which relates average limits and expected values of stationary distributions, can be applied to service systems data that are collected over a finite time interval. To make inferences based on the indirect estimator of average waiting times, we propose methods for estimating confidence intervals and for adjusting estimates to reduce bias. We show our new methods are effective using simulations and data from a US bank call center.

Second, we address important issues that need to be taken into account when testing whether real arrival data can be modeled by nonhomogeneous Poisson processes (NHPPs). We apply our method to data from a US bank call center and a hospital emergency department and demonstrate that their arrivals come from NHPPs.

Lastly, we discuss an approach to standardize the Intensive Care Unit admission process, which currently lacks a well-defined criteria. Using data from nearly 200,000 hospitalizations, we discuss how we can quantify the impact of Intensive Care Unit admission on individual patient's clinical outcomes. We then use this quantified impact and a stylized model to discuss optimal admission policies. We use simulation to compare the performance of our proposed optimal policies to the current admission policy, and show that the gain can be significant.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my special appreciation and thanks to my advisors Professors Ward Whitt, Carri Chan and Marcelo Olivares, all of whom have been tremendous mentors for me. I have been very fortunate to be able to work with and learn from them; their support and encouragement have been priceless and will continue to inspire me throughout my academic career. In addition, I thank Professors Jose Blanchet and Van-Anh Truong for serving as my committee members. I also thank my medical collaborators, Dr. Gabriel Escobar at Kaiser Permanente and Dr. Won-Chul Cha at Samsung Medical Center in South Korea. Their professional knowledge and support continuously motivate me to pursue my research interests. I am grateful to Professor Avishai Mandelbaum and the Center for Service Enterprise Engineering (SEE) at the Technion for kindly providing access to their call center data for this thesis. Finally, I thank the Samsung Foundation and the National Science Foundation (NSF Grants CMMI 1066372 and 1265070) for supporting my Ph.D. study.

Getting through my dissertation required more than academic support, and I have many, many people to thank for believing in me, encouraging me, and praying for me over the past six years. I cannot begin to express my gratitude and appreciation for my friends from high school and college, at the IEOR department, and Campus Mission Church.

A special thanks to my fiance Keun Sup who is always my support and number one fan (even of my first drafts). Words cannot express how grateful I am for the endless love and support of my parents, two younger sisters and younger brother. Their love and prayers for me have sustained me thus far. Lastly, I thank God, who gives me the power to believe in my passion and pursue my dreams. I could never have finished my Ph.D. study without His guidance and love.

Chapter 1

Introduction

The goal of this research is to facilitate providing data-driven (or evidence-based) decision support for operational decision making in service systems, especially hospitals. Data-driven decision support can both lower operating costs and raise service quality. A good example is the advances in call center operations. With active research on problem areas such as forecasting, queueing, capacity planning, and agent scheduling over the past few decades, call center operations have greatly benefited from better decision support tools. See Gans et al. (2003) and Aksin et al. (2007) and references therein for reviews of the rich literature as well as opportunities for future research.

Healthcare operations is another area where lower operating costs and higher service quality can be achieved with better decision support. Huge variations in healthcare practices reveal the problem with today's healthcare system; it is not unusual for two patients with the same condition to have different care paths and for two physicians performing the same surgery to use different procedures, drugs, devices, and equipment. These large variations suggest that there is at least some misallocation of resources, and that better management will result in lower costs and higher quality of care. Realizing this, clinical practice leaders have started paying attention to standardizing their protocols and treatment processes using evidence-based decision support (Kaplan and

Porter 2011, Chen et al. 2013). Excessive healthcare costs also point to the need for better decision support. The U.S. spent about 17% of Gross Domestic Product —$2.8 trillion or an average of $8,915 per person— on healthcare in 2012, which is two-and-a-half times more than most developed nations in the world (OECD 2014). Consequently, researchers are paying increasing attention to developing decision support tools for healthcare systems; e.g., see Armony et al. (2011) and Shi et al. (2012) and references therein.

In the following subsections we discuss the two streams of research in this dissertation. The first stream is on data-driven methods conducted with Professor Ward Whitt. The second stream is on econometric approaches to provide new policy recommendations conducted with Professors Carri Chan and Marcelo Olivares and Dr. Gabriel Escobar. The first stream is based on the three completed papers Kim and Whitt (2013e), Kim and Whitt (2014b) and Kim and Whitt (2014a), while the second stream is based on the completed paper Kim et al. (2014).

1.1 Data-Driven Methods

The Operations Research community has made extraordinary advances in building effective and sophisticated models and methods that can be used to build decision support tools for service systems. However, there is a large gap between those methods and models and the reality that is represented by service systems data. One stream of this research aims to address this problem by developing data-driven methods and models for successful design, analysis, and management, by connecting a theoretical framework with real world data and applications.

The first research topic, treated in the papers Kim and Whitt (2013e) and Kim and Whitt (2013c), is to investigate how to apply the Little's Law with data. Little's Law relates average limits and expected values of stationary distributions, but in practice we often consider averages

from data over a finite time interval. In Chapter 2, we discuss taking a statistical approach with service systems data to make inferences based on the indirect estimator of average waiting times. For stationary intervals, we advocate estimating confidence intervals based on the Central Limit Theorem version of Little's Law. For nonstationary intervals, we propose adjusted estimates to reduce bias, and advocate using data from multiple days to construct confidence intervals. We show our new methods are effective using simulations and data from a US bank call center.

Kim and Whitt (2013c) is a sequel to Chapter 2, and is not discussed in this thesis. In that sequel, we directly apply the time-varying Little's Law (TVLL), as in Bertsimas and Mourtzinou (1997) and Fralix and Riano (2010), to estimate average waiting times. We also develop useful variants of the TVLL estimator by fitting a linear or a quadratic function to arrival data. We again show that our new methods are effective using simulations and data from a US bank call center.

The second research topic, treated in the papers Kim and Whitt (2014b) and Kim and Whitt (2014a), is to examine statistical tests one can apply to confirm whether the service systems data support the candidate model. Service systems such as call centers and hospitals typically have strongly time-varying arrivals, and a natural model for their arrival process for performance analysis (e.g., in a queueing model) is a nonhomogeneous Poisson process (NHPP). Since this is such a common modeling approach, it is important to perform statistical tests with data to confirm that an NHPP is actually appropriate. We compare alternative methods that test the NHPP assumption, with a focus on examining the tests' power. We find that a careful data transformation significantly improves the power of tests. We show that a widely used data transformation by Brown et al. (2005) to test call center and hospital arrival data has a great power, but also find that a different, lesser-known data transformation by Lewis (1965) consistently provides more power.

Chapter 3, which is based on Kim and Whitt (2014a), addresses the important issues that need to be taken into account when applying the new statistical tests to real data. We find that the following three common features of arrival data are important to take into account: (1) data rounding,

e.g., to seconds, (2) over-dispersion caused by combining data from multiple days with different arrival rates, and (3) choosing subintervals over which the rate varies too much. After carefully dealing with these three features, we apply our method to data from a US bank call center and a Korean hospital emergency department. We demonstrate that their arrivals come from NHPPs (i.e., we fail to reject the null hypothesis that their arrival processes are NHPPs), but that we would have otherwise concluded that they do not come from NHPPs in general if we failed to take into account the three common features.

Kim and Whitt (2013d) is an extension of Kim and Whitt (2014b) in which we develop new methods that can be used to test whether service times can be regarded as a sequence of independent and identically distributed random variables with a specified distribution. We show that the new tests developed to test for a NHPP can be applied to this setting as well in order to increase the statistical power.

1.2 Providing Policy Recommendations

In order to improve current hospital practices using better decision support, we first need to understand critical factors that most influence the decision process of doctors and nurses as well as how different practices impact patient outcomes. This stream of research analyzes patient flow data using econometric methods to understand current practices and their impact, and leverage the findings to provide policy recommendations for improving hospital performance.

In Chapter 4, we discuss our approach to establish a data-driven decision support for Intensive Care Unit (ICU) admission. An obvious strategy for ICU admission is to admit very sick and unstable patients. However, determining which patients are the most unstable is a complex task; a systematic criteria for ICU care is currently lacking and recently, the medical community has pointed to a need to develop such criteria (Chen et al. 2013). We show, using data of nearly

200,000 hospitalizations, that although medical necessity plays a key role, operational factors such as ICU occupancy also determines which patients receive ICU care. We analyze a stylized model for ICU admission control and use simulation to evaluate the performance of alternative admission policies. By simulating a hospital with 21 ICU beds, we show that we could save about 1.9 million dollars per year by using our optimal objective policy designed to reduce readmissions and hospital length-of-stay. We believe our work is an important step in establishing an evidence-based decision support for providing ICU care.

Active collaboration with medical professionals has been critical in addressing this research. The collaborative research of Professor Carri Chan and Kaiser Permanente —the largest integrated healthcare consortium in the US— has had a significant impact on this research since 2010. The study framework for the research discussed in Chapter 4 was verified in the summer of 2011 at Kaiser Permanente hospitals where clinicians involved in making ICU admission decisions were observed and interviewed. In addition, we are currently working towards deploying an Excel-based decision support tool which would provide physicians at the Kaiser Permanente hospitals with estimates of patient outcomes based on whether s/he is admitted to the ICU or not.

A connection with Samsung Medical Center —one of the largest teaching hospitals in South Korea— was established in 2012. This connection has resulted in the research presented in Chapter 3 of this thesis and multiple ongoing projects with the physicians of Samsung Medical Center and Professor Ward Whitt.

Chapter 2

Statistical Analysis with Little's Law

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. This chapter is an edited version of Kim and Whitt (2013e).

2.1 Introduction

We have just celebrated the 50^{th} 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 Jr. (1999) and the stationary framework involving the Palm transformation by Sigman (1995) and Baccelli and Brémaud (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 §2.2. 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.

2.1.1 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 $\bar{W}(t)$ (defined in (2.1) 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 $0 \leq s \leq t$. 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 chapter we focus on the typical application in the paragraph above, 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 (2.1) 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 §2.2 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 $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$ 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. While changing the definitions is one option, we advocate *not* doing so, because it leads to problems with interpretation.

2.1.2 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. Since 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 (2010).

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 $\bar{W}(t)$ 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 $\bar{W}(t)$, 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 (2010) 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 $\bar{W}_{L,\lambda}(t) \equiv \bar{L}(t)/\bar{\lambda}(t)$ suggested by Little's law. We show the special definition used to obtain equality for $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$ within each subinterval seriously distorts the batch-means estimators when the modified definition is used within each subinterval.

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 with-

out 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 §2.2.1 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 $(\bar{L}(t), \bar{\lambda}(t), \bar{W}(t))$ as a random vector with an associated vector of finite mean values $(E[\bar{L}(t)], E[\bar{\lambda}(t)], E[\bar{W}(t)])$. 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.

Validation by Simulation

Since 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 §2.3.2.

2.1.3 Organization

Here is how the rest of this chapter is organized: In §2.2 we discuss the finite-time version of $L = \lambda W$, emphasizing the interval edge effects. In §2.3 we apply the statistical approach to a banking call center example and associated simulation models. In §2.4 we study ways to esti-

mate confidence intervals. In §2.5 we study ways to estimate and reduce the bias in the estimator $\bar{W}_{L,\lambda}(t) \equiv \bar{L}(t)/\bar{\lambda}(t)$. In §2.6 we perform experiments combining the insights in §§2.4 and 2.5; we estimate confidence intervals for refined estimators designed to reduce bias. Finally, in §2.7 we draw conclusions. Additional material appears in an appendix (Kim and Whitt 2012a) and a technical report (Kim and Whitt 2012b); the contents of both are described at the beginning of Kim and Whitt (2012a). Kim and Whitt (2013c) is a sequel to this research on estimating waiting times with the time-varying Little's law.

2.2 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).

2.2.1 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 $-\infty < A_k < D_k < \infty$, $[0, t] \cap [A_k, D_k] \neq \emptyset$ 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, $A(t) = \max\{k \geq 0 : A_k \leq t\} - R(0)$,

$t \geq 0$, 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

$$\bar{\lambda}(t) \equiv t^{-1}A(t), \quad \bar{L}(t) \equiv t^{-1}\int_0^t L(s)\,ds, \quad \bar{W}(t) \equiv (1/A(t))\sum_{k=R(0)+1}^{R(0)+A(t)} W_k, \tag{2.1}$$

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

We will focus on these averages over $[0, t]$ in (2.1), 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 n^{th} new arrival, i.e., $T_n \equiv A_{n+R(0)}$, $n \geq 0$,

$$\bar{\lambda}_n \equiv n/T_n, \quad \bar{L}_n \equiv (1/T_n)\int_0^{T_n} L(s)\,ds, \quad \bar{W}_n \equiv n^{-1}\sum_{k=R(0)+1}^{R(0)+n} W_k. \tag{2.2}$$

As in (2.1), the first two averages in (2.2) are time averages, but over the time interval $[0, T_n]$, while the last, $\bar{W}_n$, 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 (2.2) can be expressed directly in terms of the averages in (2.1): $\bar{\lambda}_n = \bar{\lambda}(T_n)$, $\bar{L}_n = \bar{L}(T_n)$ and $\bar{W}_n = \bar{W}(T_n)$, so that conclusions for (2.1) yield analogs for (2.2).

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 can we use any two of the three estimators in (2.1) to create a new alternative estimator, exploiting $L = \lambda W$:

$$\bar{L}_{W,\lambda}(t) \equiv \bar{\lambda}(t)\bar{W}(t), \quad \bar{\lambda}_{L,W}(t) \equiv \frac{\bar{L}(t)}{\bar{W}(t)} \quad \text{and} \quad \bar{W}_{L,\lambda}(t) \equiv \frac{\bar{L}(t)}{\bar{\lambda}(t)}. \tag{2.3}$$

For the typical application mentioned in §2.1 in which we observe $L(s)$, $0 \leq s \leq t$, 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 $\bar{W}_{L,\lambda}(t)$ in (2.3) as a substitute for $\bar{W}(t)$ in (2.1).

2.2.2 How the Averages in (2.1) Are Related

Figures 2.1 and 2.2 below show how the three averages in (2.1) are related. These averages are related via $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$ if the system starts and ends empty, i.e., if $R(0) = L(t) = 0$, as we show in Theorem 2.1 below. However, more generally, these averages are not simply related. To illustrate, in Figures 2.1 and 2.2 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 $0 \le s \le t$, 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 2.1 and 2.2, 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 2.2.

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 2.2. 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 2.2. 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 2.2 as the special case in which all regions except region D are empty.

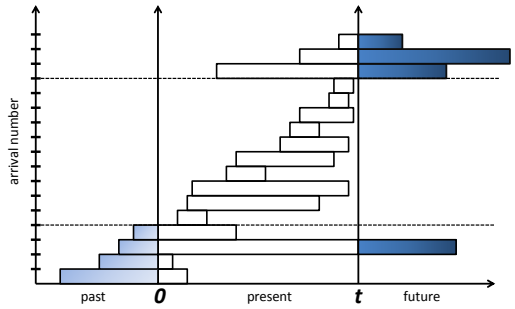


Figure 2.1: The total work in the system during the interval $[0, t]$ with edge effects: including arrivals before time 0 and departures after time t.

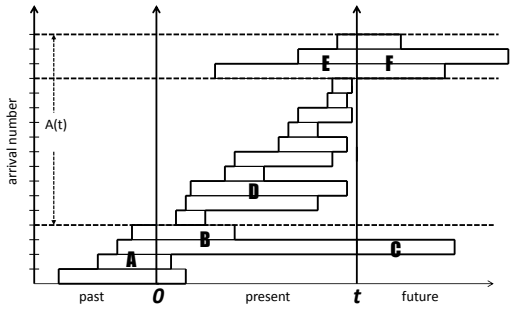


Figure 2.2: Six regions: waiting times (i) of customers that both arrive and depart inside $[0, t]$ (D), (ii) of arrivals before time 0 ($A \cup B \cup C$) and (iii) of departures after time t ($C \cup E \cup F$).

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

$$C_L(t) \equiv \int_0^t L(s)\, ds \quad \text{and} \quad C_W(t) \equiv \sum_{k=R(0)+1}^{R(0)+A(t)} W_k, \quad t \geq 0. \tag{2.4}$$

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.,

$$T_W^{(r)}(t) \equiv \sum_{k=1}^{L(t)} W_k^{r,t}, \tag{2.5}$$

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 (2.1) are the *time average* $\bar{L}(t) \equiv t^{-1} C_L(t)$ and the *customer average* $\bar{W}(t) \equiv C_W(t)/A(t)$. For a region A in Figure 2.2, let $|A|$ be the area of A. In general, the cumulative processes can be expressed in terms of the regions in Figure 2.2 as $C_L(t) = |B \cup D \cup E|$ and $C_W(t) = |D \cup E \cup F|$, while $T_W^{(r)}(0) = |B \cup C|$ and

$T_W^{(r)}(t) = |C \cup F|$, so that

$$C_L(t) - C_W(t) = |B| - |F| = |B \cup C| - |F \cup C| = T_W^{(r)}(0) - T_W^{(r)}(t). \tag{2.6}$$

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 k^{th} 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

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

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 2.1 (*traditional finite-time Little's law*) *If* $R(0) = L(t) = 0$, *then* $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$.

Proof. Under the condition, $\bar{L}(t) \equiv \frac{C_L(t)}{t} = \frac{C_W(t)}{t} = \left(\frac{A(t)}{t}\right)\left(\frac{C_W(t)}{A(t)}\right) \equiv \bar{\lambda}(t)\bar{W}(t)$. ∎

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 2.1 and 2.2, 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 Theorem 1 of Glynn and Whitt (1986). A variant appears on p. 17.4 of Mandelbaum (2010), who credits it to his student Abir Koren and emphasizes its importance for looking at data.

Theorem 2.2 (*extended finite-time Little's law*) *The averages in* (2.1) *and* (2.3) *are related by*

$$\begin{aligned}
\Delta_L(t) &\equiv \bar{L}_{W,\lambda}(t) - \bar{L}(t) = \frac{|F| - |B|}{t} = \frac{T_W^{(r)}(t) - T_W^{(r)}(0)}{t}, \\
\Delta_W(t) &\equiv \bar{W}_{L,\lambda}(t) - \bar{W}(t) = \frac{|B| - |F|}{A(t)} = -\frac{\Delta_L(t)}{\bar{\lambda}(t)} = \frac{T_W^{(r)}(0) - T_W^{(r)}(t)}{A(t)}, \\
\Delta_\lambda(t) &\equiv \bar{\lambda}_{L,W}(t) - \bar{\lambda}(t) = \left(\frac{|B| - |F|}{|D| + |E| + |F|}\right)\bar{\lambda}(t) = -\frac{\Delta_L(t)}{\bar{W}(t)},
\end{aligned} \tag{2.7}$$

where $|B|$ *is the area of the region* B *in Figure* 2.2 *and* $T_W^{(r)}(t)$ *is defined in* (2.5).

Since we focus on inferences about the average wait based on $\bar{L}(t)$ and $\bar{\lambda}(t)$ using $\bar{W}_{L,\lambda}(t)$, we focus on $\Delta_W(t)$ in (2.7). Given the customers need not depart in the order they arrive and we only observe $L(s)$, $0 \le s \le t$, the random variables $T_W^{(r)}(0)$ and $T_W^{(r)}(t)$ appearing in $\Delta_W(t)$ in (2.7) are not directly observable; we only have partial information about these random variables.

2.2.3 Alternative Definitions to Force Equality: The Inside View

Denning and Buzen (1978), Little (2011) and others have observed that we can preserve the relation $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$ in Theorem 2.1 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

$$A^{(i)}(t) \equiv R(0) + A(t), \quad t \ge 0, \quad \text{and} \quad W_k^{(i)} \equiv (D_k \wedge t) - (A_k \vee 0), \quad k \ge 1, \tag{2.8}$$

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

$$\bar{\lambda}^{(i)}(t) \equiv t^{-1}A^{(i)}(t) \quad \text{and} \quad \bar{W}^{(i)}(t) \equiv \frac{\sum_{k=1}^{A^{(i)}(t)} W_k^{(i)}}{A^{(i)}(t)}. \tag{2.9}$$

By an elementary modification of the proof of Theorem 2.1, we obtain the following "operational analysis" relation. (The equality relation corresponds to the operational Little's law on p. 235 of Denning and Buzen (1978) and Theorem LL.2 of Little (2011).)

Theorem 2.3 (*finite-time version of Little's law with altered definitions*) *With the new definitions in* (2.8) *and* (2.9), $\bar{\lambda}^{(i)}(t) \geq \bar{\lambda}(t)$, $\bar{W}^{(i)}(t) \leq \bar{W}(t)$ *and* $\bar{L}(t) = \bar{\lambda}^{(i)}(t)\bar{W}^{(i)}(t)$.

Given that we only see inside the interval $[0, t]$, the reduced waiting times are *censored*. Indeed, there is no valid upper bound on $\bar{W}(t)$ 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 $\bar{W}(t)$ 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 (2.8) also cause problems with the method of batch means used to construct confidence intervals; see §2.3.4.

2.3 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 $900 - 1200$ 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.

Since 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 chapter, 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' web sites. Each working day covers a 17-hour period from 6 am to 11 pm, referred to as $[6, 23]$.

2.3.1 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 5749 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 2.3 and 2.4 show plots of the total number of arrivals into the queue (system), $A_q(s)$, and into service, $A_{ser}(s)$, together with the total number of departures from the queue (system), $D_q(s)$, and from service, $D_{ser}(s)$, all over the interval $[0, s]$, $0 \le s \le t$, first over the entire working day $[6, 23]$ and then over the hour $[14, 15]$. These are based on the counts over 1-second subintervals. Note that the four curves in Figure 2.3 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 2.4.

From the first plot in Figure 2.3, 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

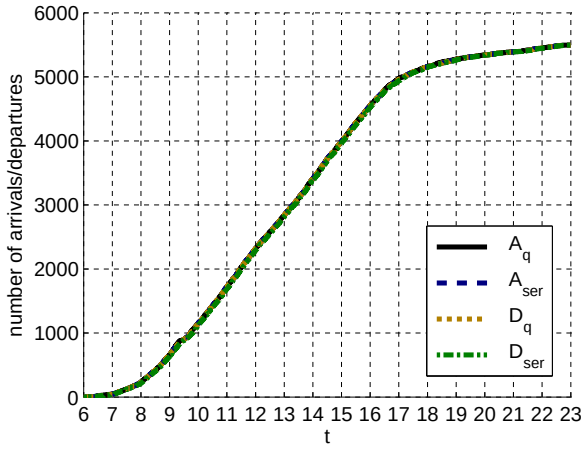


Figure 2.3: Arrivals and departures over the full day of May 25, 2001.

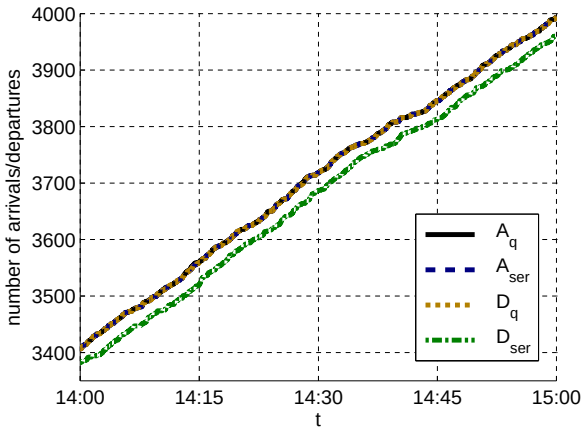


Figure 2.4: Arrivals and departures over the hour [14,15] within that day.

constant, so is the departure rate. The stationary-and-independent-increments property associated with a 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 on p. 312 of Brockwell and Davis (1991); the details appear in §2 of Kim and Whitt (2012a).

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 2.5 and 2.6. Consistent with the plots in Figure 2.3, 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 2.6 shows that the hourly averages of the waiting times do not change much, especially in the interval $[10, 16]$. During that 6-hour period $[10, 16]$, during which the system is approximately stationary, agents handled 3427 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 $[14, 15]$, 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 in using the estimates.

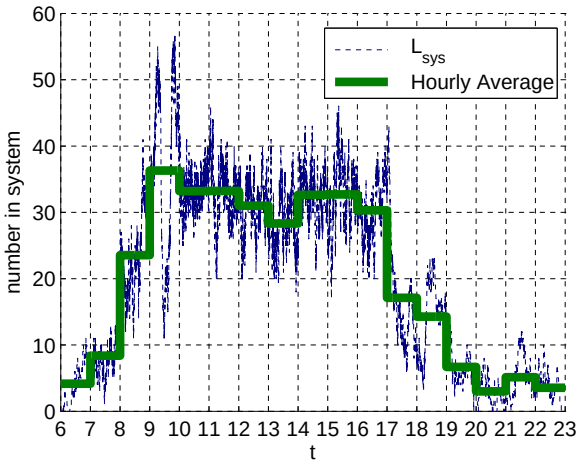


Figure 2.5: L_{sys} and its hourly averages over the full day.

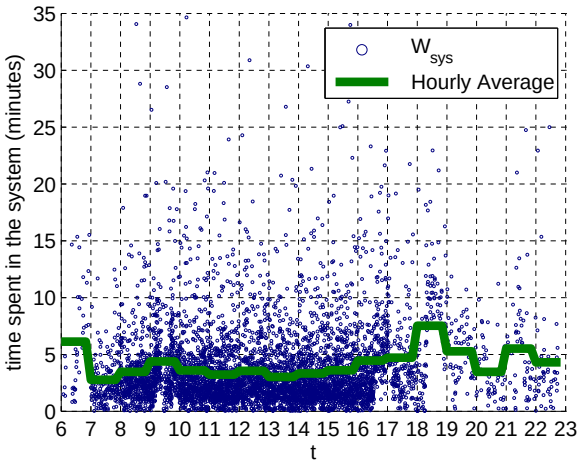


Figure 2.6: W_{sys} and its hourly averages over the full day.

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)

$$\bar{L}_{full} = 20.2 \pm 6.1, \quad \bar{\lambda}_{full} = 5.39 \pm 1.84 \quad \text{and} \quad \bar{W}_{full} = 4.18 \pm 0.56; \tag{2.10}$$

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

$$\bar{L}_{stat} = 31.8 \pm 1.0, \quad \bar{\lambda}_{stat} = 9.44 \pm 0.31 \quad \text{and} \quad \bar{W}_{stat} = 3.39 \pm 0.15 \tag{2.11}$$

For each estimate in (2.10) and (2.11), we also include the halfwidth of the 95% confidence interval, estimated as described in the §2.4.3. 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).

2.3.2 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. Since 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 $L(s)$, $0 \le s \le t$, but we do not have data on the staffing and the complex call routing. Thus, as suggested in §2.1.2, 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 Kim and Whitt (2012a), 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 Kim and Whitt (2012a)). 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. (1993a); 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 Kim and Whitt (2012a)). We simulated 1000 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.

2.3.3 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 (2.1) plus indirect estimate $\bar{W}_{L,\lambda}(t)$ from (2.3) 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 $\bar{W}_{L,\lambda}(t)$ from (2.3) 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 §2.3.2 and explained in detail in §3 of Kim and Whitt (2012a). The estimation results are shown in Table 2.1. In Table 2.1, we show direct estimates of L, λ and W from (2.1) as well as indirect estimate $\bar{W}_{L,\lambda}(t)$ from (2.3) 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. For idealized simulation models, we consider the $M_t/M/\infty$ and $M_t/M/s_t$ models with piece-wise 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. The table also shows the estimated confidence interval coverage for the two waiting time estimates for the simulations based on 1000 replications. Additional results with more values of m appear in Tables 4-9 of Kim and Whitt (2012b).

Table 2.1: Direct estimates of L, λ and W plus indirect estimate $\bar{W}_{L,\lambda}(t)$ for the time interval $[10, 16]$

case	m	$\bar{L}^{(m)}(t)$	$\bar{\lambda}^{(m)}(t)$	$\bar{W}^{(m)}(t)$	cov.	$\bar{W}^{(m)}_{L,\lambda}(t)$	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	

For large QoS parameter β, e.g., $\beta \geq 2.0$, the performance in the finite-server model is essentially the same as in the associated IS model, as can be seen from Table 2.1. 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 system increases from 31.5 to 31.6 and 32.1 for these same cases. Of special interest is the confidence interval coverage in the simulations based on 1000 replications. Table 2.1 shows it is excellent for all values of m, being very close

to the target 95.0%, for all $\beta \geq 1.5$. 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 2.1 shows that the indirect estimator $\bar{W}^{(m)}_{L,\lambda}(t)$ behaves very much the same as the direct estimator $\bar{W}(t)$. Indeed, that is consistent with the theory and other experiments in this chapter.

To illustrate what happens with a shorter sample path segment, we consider the interval $[14, 15]$. Table 2.2 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 (2012b). 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 is about 3.4 minutes, we regard only $m = 5$ appropriate for the interval $[14, 15]$ (We also note that the difference between $\bar{W}^{(m)}_{L,\lambda}(t)$ and $\bar{W}(t)$ in Table 2.2 becomes greater as m increases. See Section 2.3.4 for more discussion on this).

Table 2.2: Direct estimates of L, λ and W plus indirect estimate $\bar{W}_{L,\lambda}(t)$ for the time interval $[14, 15]$

case	m	$\bar{L}(t)$	$\bar{\lambda}(t)$	$\bar{W}(t)$	cov.	$\bar{W}^{(m)}_{L,\lambda}(t)$	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	

2.3.4 Edge Effects and the Method of Batch Means

The issue of interval edge effects discussed in §2.2 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.

The Error Due to the Interval Edge Effects

Formula (2.7) shows that the difference between $\bar{W}_{L,\lambda}(t)$ and $\bar{W}(t)$ 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 §2.3, 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 2.3. In each case, the averages over subintervals correspond to batch means. 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 4 overlapping 5-hour subintervals of $[9, 17]$, from $[9, 14]$ to $[12, 17]$, for $t = 300$ minutes. (See Tables 12-14 of Kim and Whitt (2012b) for more details.)

In Table 2.3 we see that the relative error $\Delta_W^{rel}(t) \equiv \Delta_W(t)/\bar{W}_{L,\lambda}(t)$ takes the values 20.3%, 5.6% and 0.5%, respectively for $t = 3$, 30 and 300 minutes. For the regions in Figure 2.2, for $t \geq 30$ 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 2.3 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.

Additional Error from the Altered Definitions

The altered definitions in §2.2.3 become more unattractive with batch means, because the shorter intervals distort the meaning even more. The average truncated waiting times $\bar{W}_c(t)$ in (2.9) tend to be even less than the true average waiting times W, while the average augmented arrivals $\bar{\lambda}_i(t)$

Table 2.3: Comparison of the direct and indirect estimators $\bar{W}(t)$ and $\bar{W}_{L,\lambda}(t)$ for three values of t

t	3	30	300
$\bar{W}(t)$	3.32	3.80	3.44
$\bar{W}_{L,\lambda}(t)$	3.43	3.80	3.44
$\lvert\Delta_W(t)\rvert$	0.713	0.241	0.016
$\Delta_W^{rel}(t)$	20.3%	5.6%	0.5%
$\lvert U\rvert$	311	1101	9754
$\lvert A\rvert$	34.8%	9.7%	1.4%
$\lvert B\rvert$	19.5%	9.2%	1.2%
$\lvert C\rvert$	14.3%	0.0%	0.0%
$\lvert D\rvert$	3.1%	62.8%	94.9%
$\lvert E\rvert$	8.9%	9.4%	1.2%
$\lvert F\rvert$	19.4%	8.8%	1.3%
$\lvert F\rvert - \lvert B\rvert$	6.3%	4.5%	0.4%

in (2.9) 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 $\bar{W}(t) = 3.38$ minute, while the estimators using the altered definitions in (2.9) are $\bar{\lambda}_i(t) = 9.55$ and $\bar{W}_c(t) = 3.33$. With m batches, $1 \leq m \leq 20$, the estimator $\bar{\lambda}(t)$ is unchanged and the estimator $\bar{W}(t)$ 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, $\bar{W}_c(t)$ 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 $\bar{W}_c(t)$ are 18% and 15%, respectively. When confidence intervals are formed based on batch means (for non-negligible 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 (2.9).

2.4 Confidence Intervals: Theory and Methodology

We now consider how to apply the estimator $\bar{W}_{L,\lambda}(t)$ in (2.3) to estimate a confidence interval (CI) for W in a stationary setting and for $E[\bar{W}(t)]$ 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. In order 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 §2.1.2, and as illustrated in §2.3.2.

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 §2.4.3. 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 CI's 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 and Wilson (2010), Tafazzoli et al. (2011) and references therein.

2.4.1 A Ratio Estimator

In both stationary and nonstationary settings, a CI (interval estimate) for $E[\bar{W}(t)]$ without observing the waiting times can be obtained using $\bar{W}_{L,\lambda}(t)$ if we can apply the following theorem,

implementing the delta method; see §III.3 and Proposition §IV.4.1 of Asmussen and Glynn (2007) for related results.

Theorem 2.4 (*asymptotics for the ratio of low-variability positive normal random variables*) *If there is a sequence of systems indexed by* n *such that*

$$\sqrt{n}\left(\bar{L}^{(n)}(t) - L, \bar{\lambda}^{(n)}(t) - \lambda\right) \Rightarrow N(0, \Sigma) \quad \textit{in} \quad \mathbb{R}^2 \quad \textit{as} \quad n \to \infty, \tag{2.12}$$

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* $\bar{W}^{(n)}(t)$ *satisfies*

$$\bar{W}^{(n)}(t)/\bar{W}_{L,\lambda}^{(n)}(t) \Rightarrow 1 \quad \textit{as} \quad n \to \infty, \tag{2.13}$$

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

$$\sqrt{n}\left(\bar{W}^{(n)}(t) - (L/\lambda)\right) \Rightarrow N(0, \sigma_W^2) \quad \textit{in} \quad \mathbb{R} \quad \textit{as} \quad n \to \infty \tag{2.14}$$

for

$$\sigma_W^2 = \frac{1}{\lambda^2}\left(\sigma_L^2 - \frac{2L\sigma_{L,\lambda}^2}{\lambda} + \frac{L^2\sigma_\lambda^2}{\lambda^2}\right). \tag{2.15}$$

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

$$\frac{\bar{L}^{(n)}(t)}{\bar{\lambda}^{(n)}(t)} = \frac{L}{\lambda} + \frac{\bar{L}^{(n)}(t) - L}{\lambda} - \frac{L(\bar{\lambda}^{(n)}(t) - \lambda)}{\lambda^2} + o(\max\{|\bar{L}^{(n)}(t) - L|, |\bar{\lambda}^{(n)}(t) - \lambda|\}), \tag{2.16}$$

so that

$$\sqrt{n}\left(\bar{W}_{L,\lambda}^{(n)}(t) - (L/\lambda)\right) = \frac{\sqrt{n}(\bar{L}^{(n)}(t) - L)}{\lambda} - \frac{\sqrt{n}L(\bar{\lambda}^{(n)}(t) - \lambda)}{\lambda^2} + o(1) \quad \text{as} \quad n \to \infty, \tag{2.17}$$

from which (2.14) follows, given (2.12) and (2.13). ∎

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 $E[\bar{W}^{(n)}(t)]$ would then be $[\bar{W}^{(n)}_{L,\lambda}(t) - 1.96\sigma_W/\sqrt{n}, \bar{W}^{(n)}_{L,\lambda}(t) + 1.96\sigma_W/\sqrt{n}]$, where σ_W is the square root of the variance σ^2_W in (2.15). Since the variance σ^2_W in (2.15) is typically unknown, we must estimate it. That can be done by inserting estimates for all the components of (2.15). Assuming that the estimates converge as $n \to \infty$, we still have asymptotic normality with the estimated values of the variance σ^2_W.

The sequence of systems indexed by n satisfying condition (2.12) in Theorem 2.4 can arise in two natural ways: First, condition (2.12) 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, $E[\bar{W}^{(n)}(t)] \to W$ as $n \to \infty$, and we are simply estimating W. Second, whether or not there is a stationary environment, condition (2.12) 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 (2.13) in Theorem 2.4 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 (2.13) 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 §2.5.

2.4.2 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 (2.1) as estimators of the unknown parameters L, λ and W. The CLT states that, under very general regularity conditions,

$$(\hat{L}(t), \hat{\lambda}(t), \hat{W}(t), \hat{L}_{W,\lambda}(t), \hat{\lambda}_{L,W}(t), \hat{W}_{L,\lambda}(t)) \Rightarrow (X_L, X_\lambda, X_W, X_L, X_\lambda, X_W) \quad \text{in} \quad \mathbb{R}^6 \quad (2.18)$$

as $t \to \infty$, where

$$\begin{aligned} (\hat{L}(t), \hat{\lambda}(t), \hat{W}(t)) &\equiv \sqrt{t}\left(\bar{L}(t) - L, \bar{\lambda}(t) - \lambda, \bar{W}(t) - W\right), \\ (\hat{L}_{W,\lambda}(t), \hat{\lambda}_{L,W}(t), \hat{W}_{L,\lambda}(t)) &\equiv \sqrt{t}\left(\bar{L}_{W,\lambda}(t) - L, \bar{\lambda}_{L,W}(t) - \lambda, \bar{W}_{L,\lambda}(t) - W\right), \end{aligned}$$

with the averages given in (2.1) and (2.3), 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

$$\begin{aligned} \sigma_W^2 &\equiv Var(X_W) = E[X_W^2] = \lambda^{-2}(\sigma_L^2 - 2W\sigma_{\lambda,L}^2 + W^2\sigma_\lambda^2), \\ \sigma_{L,W}^2 &\equiv Cov(X_L, X_W) = E[X_L X_W] = \lambda^{-1}(\sigma_L^2 - W\sigma_{\lambda,L}^2), \\ \sigma_{W,\lambda}^2 &\equiv Cov(X_W, X_\lambda) = E[X_W X_\lambda] = \lambda^{-1}(\sigma_{\lambda,L}^2 - W\sigma_\lambda^2). \end{aligned} \quad (2.19)$$

Note that σ_W^2 in (2.19) agrees with (2.15).

Under general regularity conditions (essentially, if $t^{1/2}T_W^{(r)}(t) \Rightarrow 0$ for $T_W^{(r)}(t)$ in (2.5)), a functional central limit theorem (FCLT) generalization of the joint CLT in (2.18) is valid if a FCLT is valid in $\mathbb{R}^2$ for any two of the first three components. For example, it suffices start with the

(FCLT generalization of) the bivariate CLT

$$\sqrt{t}(\bar{L}(t) - L, \bar{\lambda}(t) - \lambda) \Rightarrow (X_L, X_\lambda) \quad \text{in} \quad \mathbb{R}^2 \quad \text{as} \quad t \to \infty, \tag{2.20}$$

where the limit (X_L, X_λ) is a bivariate mean-zero Gaussian random vector with variances σ^2_λ, σ^2_L and covariance $\sigma^2_{\lambda,L}$. Natural sufficient conditions are based on regenerative structure for the stochastic process $\{L(t) : t \geq 0\}$, as in §VI.3 of Asmussen (2003) and Glynn and Whitt (1987). We directly assume that the limit in (2.18) is valid, and discuss how to apply it. Note that condition (2.20) coincides with condition (2.12) in Theorem 2.4, but now the conclusion directly gives a CLT for $\bar{W}(t)$ as well as for $\bar{W}_{L,\lambda}(t)$.

The form of the limit in (2.18) implies that the alternative estimators $\bar{L}_{W,\lambda}(t)$, $\bar{\lambda}_{L,W}(t)$ and $\bar{W}_{L,\lambda}(t)$ in (2.3) not only converge to the same limits L, λ and W just as the natural estimators $\bar{L}(t)$, $\bar{\lambda}(t)$ and $\bar{W}(t)$ in (2.1) do, but also the corresponding CLT-scaled random variables are asymptotically equivalent as well, i.e.,

$$\|(\hat{L}(t), \hat{\lambda}(t), \hat{W}(t)) - (\hat{L}_{W,\lambda}(t), \hat{\lambda}_{L,W}(t), \hat{W}_{L,\lambda}(t))\| \Rightarrow 0 \quad \text{as} \quad t \to \infty,$$

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 (2.3) as it is for the natural estimators in (2.1) (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.

2.4.3 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 $\{(A(s), L(s)) : 0 \le s \le t\}$ and $\{W_k : R(0) + 1 \le k \le R(0) + A(t)\}$ 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 $[(k-1)t/m, kt/m]$, $1 \le k \le m$. 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 $\bar{W}_k(t, m)$ denote the averages over the interval $[(k-1)t/m, kt/m]$, i.e.,

$$\begin{gathered}
\bar{A}_k(t,m) \equiv (m/t)A_k(t,m), \quad \bar{L}_k(t,m) \equiv (m/t)L_k(t,m), \\
\bar{W}_k(t,m) \equiv (1/A_k(t,m))W_k(t,m), \\
A_k(t,m) \equiv A(kt/m) - A((k-1)t/m), \quad L_k(t,m) \equiv \int_{(k-1)t/m}^{kt/m} L(s)\, ds, \\
W_k(t,m) \equiv \sum_{j=R_{k-1}+1}^{R_{k-1}+A_k(t,m)} W_j.
\end{gathered} \tag{2.21}$$

The FCLT version of the CLT in the previous section implies that, as $t \to \infty$, the vector of scaled batch means $\sqrt{t/m}(\bar{A}_k(t,m) - \lambda, \bar{L}_k(t,m) - L, \bar{W}_k(t,m) - W)$, $1 \le k \le m$, 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 2.4, as $t \to \infty$, the associated scaled vector $\sqrt{t/m}(\bar{W}_{L,\lambda,k}(t,m) - W)$, $1 \le k \le m$, are asymptotically m i.i.d. mean-zero random variables with variance σ_W^2 in (2.15). Hence, as $t \to \infty$, also

$$\frac{\sum_{k=1}^m (\bar{W}_{L,\lambda,k}(t,m) - \bar{W}_{L,\lambda}^{(m)}(t))}{\sqrt{S_{(m)}^2(t)/m}} \Rightarrow t_{m-1}, \tag{2.22}$$

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

$$\bar{W}_{L,\lambda}^{(m)}(t) \equiv \frac{1}{m}\sum_{k=1}^{m}\bar{W}_{L,\lambda,k}(t,m) \quad \text{and} \quad S_{(m)}^2(t) \equiv \frac{1}{m-1}\sum_{k=1}^{m}(\bar{W}_{L,\lambda,k}(t,m) - \bar{W}_{L,\lambda}^{(m)}(t))^2. \quad (2.23)$$

Thus, $[\bar{W}_{L,\lambda}^{(m)}(t) - t_{0.025,m-1}S_{(m)}(t)/\sqrt{m}, \bar{W}_{L,\lambda}^{(m)}(t) + t_{0.025,m-1}S_{(m)}(t)/\sqrt{m}]$ is an approximate 95% confidence interval for W based on the t distribution and the average $\bar{W}_{L,\lambda}^{(m)}(t)$ 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.

2.5 Estimating and Reducing the Bias

We now discuss ways to estimate and reduce the bias in the estimator $\bar{W}_{L,\lambda}(t)$ in (2.3) as an estimator for $E[\bar{W}(t)]$ for $\bar{W}(t)$ in (2.1). In doing so, we are primarily concerned with nonstationary settings. In stationary settings, $\bar{W}(t)$ in (2.1) is typically a biased estimator of W, while $\bar{W}_{L,\lambda}(t)$ is typically a biased estimator of both W and $E[\bar{W}(t)]$, but these biases are less likely to be serious, e.g., see §2.5.4.

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 §2.5.4, we show that the bias in $\bar{W}(t)$ for estimating W tends to be negligible in the infinite-server paradigm.

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

Since the bias in $\bar{W}_{L,\lambda}(t)$ as an estimator for $E[\bar{W}(t)]$ is $E[\Delta_W(t)]$ for $\Delta_W(t) \equiv \bar{W}_{L,\lambda}(t) - \bar{W}(t)$ in (2.7), we can apply Theorem 2.2 to obtain an exact expression for the bias $E[\Delta_W(t)]$. We also give the conditional bias $E[\Delta_W(t)|\mathcal{O}_t]$ 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 2.1 (*exact bias and conditional bias*) *The bias in* $\bar{W}_{L,\lambda}(t)$ *in* (2.3) *as an estimator for* $E[\bar{W}(t)]$ *for* $\bar{W}(t)$ *in* (2.1) *is* $E[\Delta_W(t)] = E[E[\Delta_W(t)|\mathcal{O}(t)]]$, *where* $\Delta_W(t)$ *is given in* (2.7), *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*

$$E[\Delta_W(t)|\mathcal{O}_t] = \frac{\sum_{k=1}^{R(0)} E[W_k^{r,0}|\mathcal{O}_t] - \sum_{k=1}^{L(t)} E[W_k^{r,t}|\mathcal{O}_t]}{A(t)}. \tag{2.24}$$

Proof. Apply Theorem 2.2 using (2.5). ∎

2.5.2 Two Approximations

The bias in Corollary 2.1 is not easy to analyze. Given that $(R(0), L(t), A(t))$ is observed, it remains to estimate the conditional residual waiting times $E[W_k^{r,0}|\mathcal{O}_t]$, $1 \le k \le R(0)$, and $E[W_k^{r,t}|\mathcal{O}_t]$, $1 \le k \le L(t)$. The conditional expectations $E[W_k^{r,0}|\mathcal{O}_t]$ 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.

Simplification from the Bias Approximation Assumption

As t increases, we expect the "initial edge effect" $\{R(0), W_k^{r,0}; 1 \le k \le R(0)\}$ to be approximately independent of the "terminal edge effect" $\{L(t), W_k^{r,t}; 1 \le k \le L(t)\}$ 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 $\mathcal{O}(t) \equiv (t, \bar{L}(t), \bar{\lambda}(t), R(0), L(t)), t \geq 0$,

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

Invoking the BAA, we obtain the following approximation directly from Corollary 2.1:

$$E[\Delta_W(t)|\mathcal{O}_t] \approx \frac{\sum_{k=1}^{R(0)} E[W_k^{r,0}|R(0)] - \sum_{k=1}^{L(t)} E[W_k^{r,t}|L(t)]}{A(t)}. \tag{2.25}$$

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), $\{R(0), W_k^{r,0}; 1 \leq k \leq R(0)\}$ will be asymptotically independent of $\{L(t), W_k^{r,t}; 1 \leq k \leq L(t)\}$.

Using $\bar{W}_{L,\lambda}(t)$ to Estimate the Residual Waiting Times

We can obtain an applicable estimate of the conditional bias $E[\Delta_W(t)|\mathcal{O}_t]$ in (2.24) if we estimate all the remaining conditional waiting times by the observed $\bar{W}_{L,\lambda}(t)$. 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

$$E[\Delta_W(t)|\mathcal{O}_t] \approx \frac{(R(0) - L(t))\bar{W}_{L,\lambda}(t)}{A(t)} \quad \text{for} \quad \mathcal{O}(t) \equiv (R(0), L(t), \bar{L}(t), \bar{\lambda}(t)). \tag{2.26}$$

We can apply approximation (2.26) to obtain the new candidate *refined estimator* of $E[\bar{W}(t)]$,

exploiting the observed vector $(R(0), L(t), A(t))$:

$$\bar{W}_{L,\lambda,r}(t) \equiv \bar{W}_{L,\lambda}(t) - E[\Delta_W(t)|\mathcal{O}_t] \approx \bar{W}_{L,\lambda}(t)\left(1 - \frac{R(0) - L(t)}{A(t)}\right). \tag{2.27}$$

(The refined estimator $\bar{W}_{L,\lambda,r}(t)$ in (2.27) is a candidate refinement of the indirect estimator $\bar{W}_{L,\lambda}(t)$ (2.3).) The associated approximate relative conditional bias is thus

$$E[\Delta_W^{rel}(t)|\mathcal{O}(t)] \equiv \frac{E[\Delta_W(t)|\mathcal{O}_t]}{E[\bar{W}(t)]} \approx \frac{E[\Delta_W(t)|\mathcal{O}_t]}{\bar{W}_{L,\lambda}(t)} \approx \frac{R(0) - L(t)}{A(t)}. \tag{2.28}$$

In the next section we show that the analysis in (2.26)-(2.28) can be supported theoretically in the infinite-server paradigm when the waiting times are exponential, so we propose the refined estimator in (2.27) 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 §2.5.5.

2.5.3 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 (2.25) becomes

$$E[\Delta_W(t)|\mathcal{O}_t] \approx (R(0) - L(t))E[S]/A(t). \tag{2.29}$$

Since the waiting times coincide with the service times in the IS model, it is natural to use the observed $\bar{W}_{L,\lambda}(t)$ as an initial estimate of $E[S]$. If we use $\bar{W}_{L,\lambda}(t)$ as an estimate of $E[S]$ in (2.29), then the formula in (2.29) reduces to the bias approximation in (2.26). Thus, under these approximations, the refined estimator (2.27) becomes unbiased. Hence, we propose the refined estimator in (2.27) for light-to-moderately-loaded many-server systems with service time distributions not

too far from exponential.

To better understand the consequence of non-exponential service times in the infinite-server paradigm, we now consider the $M_t/GI/\infty$ IS model with non-exponential 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 $G(x) \equiv P(S \le x)$ with $E[S^2] < \infty$ and thus finite squared coefficient of variation (SCV) $c_S^2 \equiv Var(S)/E[S]^2$. 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,

$$P(S_e \le x) \equiv \frac{1}{E[S]} \int_0^x G^c(u)\, du \quad \text{and} \quad E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}. \tag{2.30}$$

For this IS model, we can characterize the conditional expected value of the remaining work $T_W^{(r)}(t)$ in (2.5) and (2.7) given $L(t)$, but it requires the full waiting-time cdf G.

Theorem 2.5 (*total remaining work for the $M_t/GI/\infty$ infinite-server model*) *For the $M_t/GI/\infty$ model above,*

$$E[T_W^{(r)}(t)|L(t)] = \frac{L(t) \int_0^\infty \lambda(t-u) E[S-u; S>u]\, du}{E[L(t)]}, \tag{2.31}$$

for $T_W^{(r)}(t)$ in (2.5), where $E[S-u; S>u] = E[S-u|S>u]P(S>u)$, $E[S-u|S>u] = \int_0^\infty (G^c(x-u)/G^c(u))\, dx$, and

$$E[L(t)] = \int_0^\infty \lambda(t-u) G^c(u)\, du = E[\lambda(t-S_e)]E[S], \quad t \ge 0. \tag{2.32}$$

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

$$P(S_t > x) = \frac{\int_0^\infty \lambda(t-u) P(S > x+u)\, du}{E[L(t)]}, \tag{2.33}$$

for $E[L(t)]$ given in (2.32), by Theorem 2.1 of Goldberg and Whitt (2008), which draws on Eick et al. (1993b). 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 (2.32) is given in Theorem 1 of Eick et al. (1993b). ∎

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

$$E[\Delta_W(t)|\mathcal{O}_t] \approx \frac{E[T_W^{(r)}(0)|L(0)] - E[T_W^{(r)}(t)|L(t)]}{A(t)}, \tag{2.34}$$

where (2.31) 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 (2.34) based on Theorem 2.5 would not appear to be very useful, but we now show that it provides strong support for the refined estimator in (2.27) 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 (2.31) reduces to $E[T_W^{(r)}(t)|L(t)] = L(t)E[S]$, taking us back to (2.26). Second, for the stationary $M/GI/\infty$ model starting empty in the infinite past, S_t in (2.33) is distributed as S_e in (2.30), so that formula (2.31) reduces to $E[T_W^{(r)}(t)|L(t)] = L(t)E[S_e] = L(t)E[S](c_s^2 + 1)/2$ and (2.34) reduces to $E[\Delta_W(t)|\mathcal{O}_t] = (R(0) - L(t))E[S](c_s^2 + 1)/2A(t)$, 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 2.5. Thus, we conclude that this section provides strong support for the refined estimator $\bar{W}_{L,\lambda,r}(t)$ in (2.27) 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.

2.5.4 Bias of $\bar{W}(t)$ in the Infinite-Server Paradigm

We now observe that the bias of $\bar{W}(t)$ 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

$$E[\bar{W}(t)|A(t) > 0] = E[E[\bar{W}(t)|A(t)]|A(t) > 0] = E[W|A(t) > 0] = W. \tag{2.35}$$

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

Theorem 2.6 (*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,*

$$E[\bar{W}(t)] = WP(A(t) > 0). \tag{2.36}$$

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

2.5.5 The Single-Server Paradigm

To show that the refined estimator $\bar{W}_{L,\lambda,r}(t)$ in (2.27) 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 (2.5) 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

$$T_W^{(r)}(t) \equiv \sum_{k=1}^{L(t)} W_k^{r,t} = L(t)S^{(r)}(t) + \sum_{k=1}^{L(t)-1} (L(t)-k)S_{k+1}, \tag{2.37}$$

where S_k, $k \geq 2$, 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 2.7 (*bias reduction for the $G_t/M/1/\infty$ model*) *For the $G_t/M/1/\infty$ model,*

$$E[T_W^{(r)}(t)|L(t), E[S]] = L(t)(L(t)+1)E[S]/2. \tag{2.38}$$

so that, if $T_W^{(r)}(0)$ is fully observable in $[0,t]$, then for $\mathcal{O}(t) \equiv (L(t), \bar{L}(t), \bar{\lambda}(t), T_W^{(r)}(0), E[S])$,

$$E[\Delta_W(t)|\mathcal{O}_t] = \frac{T_W^{(r)}(0) - L(t)(L(t)+1)E[S]/2}{A(t)}. \tag{2.39}$$

Proof. Formula (2.39) follows directly from (2.38), which in turn follows from (2.37) 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 2.7 to obtain the single-server refined estimator

$$\bar{W}_{L,\lambda,r,1}(t) \equiv \bar{W}_{L,\lambda}(t) - \frac{T_W^{(r)}(0) - L(t)(L(t)+1)E[S]/2}{A(t)}. \tag{2.40}$$

Even if we do not know the mean $E[S]$, formulas (2.38)-(2.40) provide important insight, showing that $E[T_W^r(t)|L(t), E[S]]$ is approximately proportional to $L(t)^2$ instead of $L(t)$ as in (2.26) and §2.5.3. We next show that the bias in (2.39) 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 §2.5.5, 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 3 values of the constant arrival rate λ: 0.7, 1.0 and 2.0. The respective 95% confidence intervals (CI's) for the exact value of $E[\bar{W}(t)]$ estimated by the sample average of 1000 replications of $\bar{W}(t)$ 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 $\bar{W}_{L,\lambda,r,1}(t)$ in (2.40) has essentially no bias at all, without expense of wider CI's, the corresponding CI's for it based on the same 1000 replications were 1.90 ± 0.08, 2.68 ± 0.11 and 6.38 ± 0.18. In contrast, the unrefined $\bar{W}_{L,\lambda}(t)$ in (2.3) produced the corresponding tighter erroneous CI's 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 (2.27) should perform well here. That is confirmed by the corresponding CI's 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 $\lambda = 2.0$.

2.6 Confidence Intervals for the Refined Estimator

We now see how the two statistical techniques in §§2.4 and 2.5 can be combined. We estimate confidence intervals for the refined estimators in (2.27) and (2.40) as well as the other estimators in (2.1) and (2.3).

2.6.1 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 tran-

sient $M/M/1$ queue in §2.5.5. We now show how the sample average approach can be applied to estimate confidence intervals for the refined estimator in (2.40) 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 2.4 shows that the unrefined estimator $\bar{W}_{L,\lambda}(t)$ in (2.3) does a very poor job in estimating the mean wait because of the bias, but the performance of the refined estimator $\bar{W}_{L,\lambda,r}(t)$ in (2.27) and the direct estimator $\bar{W}(t)$ is not too bad. The 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. 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 non-normality and skewness; see §4 of Kim and Whitt (2012a) for details. In an effort to obtain a better estimate of confidence intervals, one can consider using 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 Kim and Whitt (2012a)). For more discussion on skewness-adjusted CI, see Johnson (1978), Willink (2005); in context of batch means and their residual skewness and correlations, see Alexopoulos and Goldsman (2004), Tafazzoli et al. (2011), Tafazzoli and Wilson (2010) and references therein.

Table 2.4: Confidence intervals for the mean wait in the transient $M/M/1$ Queue for $\lambda = 0.7, 1.0,$ and 2.0.

λ	$\bar{L}(t)$	$\bar{\lambda}(t)$	$\bar{W}(t)$	cov.	$\bar{W}_{L,\lambda}(t)$	cov.	$\bar{W}_{L,\lambda,r}(t)$	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%

2.6.2 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 (2.28) 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 §2.3 to confirm that observation and show that the refined estimator in (2.27) 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 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 Section 2.6.3). In particular, we compute the average over the 18 days of the absolute errors $|\bar{W}_{L,\lambda}(t) - \bar{W}(t)|$ (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 2.5 highlights the results; AAE and ASE of each subinterval (hours and half hours) are averaged over the intervals $[6, 10]$, $[10, 16]$, $[16, 23]$ and all day, and are again averaged over 18 weekdays in the call center example. More details appear in Tables 15 and 16 in Kim and Whitt (2012b).

Table 2.5: Comparison of the refined estimator $\bar{W}_{L,\lambda,r}(t)$ to the unrefined estimator $\bar{W}_{L,\lambda}(t)$

Subinterval Length	Intervals averaged over	unrefined in (2.3) $\bar{W}_{L,\lambda}(t)$	AAE	ASE	refined in (2.27) $\bar{W}_{L,\lambda,r}(t)$	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

Table 2.5 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 $\bar{W}(t)$ during $[6, 10]$ when the arrival rate is increasing, and that it overestimates $\bar{W}(t)$ during $[16, 23]$ when the arrival rate is decreasing, as expected.

2.6.3 Sample Averages Over Separate Days

For many service systems, whether stationary or not, we may be able to estimate CI's for $E[\bar{W}(t)]$ in (2.1) without observing the waiting times via $E[\bar{W}_{L,\lambda}(t)]$ in (2.3) 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 2.4, the associated random variable $\bar{W}_{L,\lambda}(t)$ should be approximately Gaussian as well with (unknown) variance given in (2.15). We also assume that any refinement $\bar{W}_{L,\lambda,r}(t)$ 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 $\bar{W}_{L,\lambda}(t)$ or (preferably) its refinement $\bar{W}_{L,\lambda,r}(t)$ based on the bias analysis described in §2.5 for day i. Let the sample mean and variance be

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{2.41}$$

Then $(\bar{X}_n - E[\bar{W}(t)])/\sqrt{S_n^2/n}$ 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 $E[\bar{W}(t)]$.

To assess how well indirect estimators perform in estimating $E[\bar{W}(t)]$ over separate days and in different settings, we again consider our call center data and divide each day into 3 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 $\bar{W}_{L,\lambda,r}(t)$ in (2.27) and the unrefined estimator $\bar{W}_{L,\lambda}(t)$ in (2.3), as well as that of direct estimator is illustrated in Table 2.6. (Additional estimation results appear in Tables 17-19 of Kim and Whitt (2012b).) We see that the refined estimator $\bar{W}_{L,\lambda,r}(t)$ behaves very similar 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 2.6: Estimating $E[\bar{W}(t)]$ and its associated 95% confidence interval over 18 weekdays in the call center example

Intervals	direct estimator $\bar{W}(t)$	unrefined in (2.3) $\bar{W}_{L,\lambda}(t)$	refined in (2.27) $\bar{W}_{L,\lambda,r}(t)$
$[6, 10]$	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

2.7 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 §2.2.3, it is possible to modify the definitions so that the relation $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$ always holds for finite averages, but we advocate not doing so. Instead, we advocate taking a statistical approach, estimating confidence intervals (§2.4) and considering modified estimators that reduce bias (§2.5), which exploit the extended finite-time Little's law in Theorem 2.2. We have illustrated the statistical approach by applying it to the call center example in §2.3. We have focused on the problem of estimating the unknown mean values W and $E[\bar{W}(t)]$ using $\bar{W}_{L,\lambda}(t) \equiv \bar{L}(t)/\bar{\lambda}(t)$ when the waiting times cannot be directly observed.

Chapter 3

Are Call Center and Hospital Arrivals Well Modeled by NHPPs?

Service systems such as call centers and hospitals typically have strongly time-varying arrivals. A natural model for such an arrival process is a nonhomogeneous Poisson process (NHPP), but that should be tested by applying appropriate statistical tests to arrival data. Assuming that the NHPP has a rate that can be regarded as approximately piecewise-constant, a Kolmogorov-Smirnov (KS) statistical test of a Poisson process (PP) can be applied to test for a NHPP, by combining data from separate subintervals, exploiting the classical conditional-uniform property. In this chapter we apply KS tests to banking call center and hospital emergency department arrival data and show that they are consistent with the NHPP property, but only if that data is analyzed carefully. Initial testing rejected the NHPP null hypothesis, because it failed to take account of three common features of arrival data: (i) data rounding, e.g., to seconds, (ii) choosing subintervals over which the rate varies too much, and (iii) over-dispersion caused by combining data from fixed hours on a fixed day of the week over multiple weeks that do not have the same arrival rate. In this chapter we investigate how to address each of these three problems. This chapter is an edited version of

Kim and Whitt (2014a).

3.1 Introduction

Significant effort is being made to apply operations management approaches to improve the performance of service systems such as call centers and hospitals (Aksin et al. 2007, Armony et al. 2011). Since call centers and hospitals typically have strongly time-varying arrivals, when analyzing the performance to allocate resources (e.g., staffing), it is natural to model the arrival process as a nonhomogeneous Poisson process (NHPP). We usually expect these arrival processes to be well modeled by NHPP's, because the arrivals typically arise from the independent decisions of many different people, each of whom uses the service system infrequently. Mathematical support is provided by the Poisson superposition theorem; e.g., see Barbour et al. (1992), §11.2 of Daley and Vere-Jones (2008) and §9.8 of Whitt (2002).

Nevertheless, there are phenomena that can prevent the Poisson property from occurring. For example, scheduled appointments as at a doctor's office and enforced separation in airplane landings at airports tend to make the arrival processes less variable or less bursty than Poisson. On the other hand, arrival processes tend to be more variable or more bursty than Poisson if they involve overflows from other finite-capacity systems, as occur in hospitals (Asaduzzaman et al. 2010, Litvak et al. 2008) and in requests for reservations at hotels, because the overflows tend to occur in clumps during those intervals when the first system is full. Indeed, there is a long history studying overflow systems in teletraffic engineering (Cooper 1982, Wilkinson 1956). Bursty arrival processes also occur if the arrivals directly occur in batches, as in arrivals to hospitals from accidents. In restaurants arrivals occur in groups, but the group usually can be regarded as a single customer. In contrast, in hospitals batch arrivals typically use resources as individuals. From the extensive experience in teletraffic engineering, it is known that departures from the Poisson property can have a strong impact upon performance; that is supported by recent work in Li and Whitt (2013), Pang

and Whitt (2012). We emphasize this key point by showing the results of simulation experiments in §3 of Kim and Whitt (2013b).

3.1.1 Exploiting the Conditional Uniform Property

Hence, to study the performance of any given service system, it is appropriate to look closely at arrival data and see if an NHPP is appropriate. A statistical test of an NHPP was suggested by Brown et al. (2005). Assuming that the arrival rate can be regarded as approximately piecewise-constant (PC), they proposed applying the classical *conditional uniform* (CU) property over each interval where the rate is approximately constant. For a Poisson process (PP), the CU property states that, conditional on the number n of arrivals in any interval $[0, T]$, the n ordered arrival times, each divided by T, are distributed as the order statistics of n independent and identically distributed (i.i.d.) random variables, each uniformly distributed on the interval $[0, 1]$. Thus, under the NHPP hypothesis, if we condition in that way, the arrival data over several intervals of each day and over multiple days can all be combined into one collection of i.i.d. random variables uniformly distributed over $[0, 1]$.

Brown et al. (2005) suggested applying the Kolmogorov-Smirnov (KS) statistical test to see if the resulting data is consistent with an i.i.d. sequence of uniform random variables. To test for n i.i.d. random variables X_j with *cumulative distribution function* (cdf) F, the KS statistic is the uniform distance between the *empirical cdf* (ecdf)

$$\bar{F}_n(t) \equiv \frac{1}{n} \sum_{j=1}^{n} 1_{\{(X_j/T) \le t\}}, \quad 0 \le t \le 1, \tag{3.1}$$

and the cdf F, i.e., the *KS test statistic* is

$$D_n \equiv \sup_{0 \le t \le 1} |\bar{F}_n(t) - F(t)|. \tag{3.2}$$

We call the KS test of a PP directly after applying the CU property to a PC NHPP the CU KS test; it uses (3.2) with the uniform cdf $F(t) = t$. The KS test compares the observed value of D_n to the *critical value*, $\delta(n, \alpha)$; the PP null hypothesis H_0 is rejected at significance level α if $D_n > \delta(n, \alpha)$ where $P(D_n > \delta(n, \alpha)|H_0) = \alpha$. In this chapter, we always take α to be 0.05, in which case it is known that $\delta(n, \alpha) \approx 1.36/\sqrt{n}$ for $n > 35$; see Simard and L'Ecuyer (2011) and references therein.

3.1.2 The Possibility of a Random Rate Function

It is significant that the CU property eliminates all nuisance parameters; the final representation is independent of the rate of the PP on each subinterval. That helps for testing a PC NHPP, because it allows us to combine data from separate intervals with different rates on each interval. The CU KS test is thus the same as if it were for a PP. However, it is important to recognize that the constant rate on each subinterval could be random; a good test result does *not* support any candidate rate or imply that the rate on each subinterval is deterministic. Thus those issues remain to be addressed. For dynamic time-varying estimation needed for staffing, that can present a challenging forecasting problem, as reviewed in Ibrahim et al. (2012) and references therein.

By applying the CU transformation to different days separately, as well as to different subintervals within each day as needed to warrant the PC rate approximation, this method accommodates the commonly occurring phenomenon of day-to-day variation, in which the rate of the Poisson process randomly varies over different days; see, e.g., Avramidis et al. (2004), Ibrahim et al. (2012), Jongbloed and Koole (2001). Indeed, if the CU transformation is applied in that way (by combining the data over multiple days treated separately), then the statistical test should be regarded as a test of a Cox process, i.e., for a doubly stochastic PP, where the rate is random over the days, but is constant over each subinterval over which the CU transformation is applied.

Indeed, even though we will not address that issue here, there is statistical evidence that the

rate function often should be regarded as random over successive days, even for the same day of the week. It is important to note that, where these more complex models with random rate function are used, as in Bassamboo and Zeevi (2009) and Ibrahim et al. (2012), it invariably is *assumed* that the arrival process is a Cox process, i.e., that it has the Poisson property with time-varying stochastic rate function. The statistical tests we consider can be used to test if that assumed model is appropriate.

3.1.3 An Additional Data Transformation

In fact, Brown et al. (2005) did *not* actually apply the CU KS test. Instead, they suggested applying the KS test based on the CU property *after* performing an additional logarithmic data transformation. Kim and Whitt (2014b) investigated why an additional data transformation is needed and what form it should take. They showed through large-sample asymptotic analysis and extensive simulation experiments that the CU KS test of a PP has remarkably little power against alternative processes with non-exponential interarrival-time distributions. They showed that low power occurs because the CU property focuses on the arrival times instead of the interarrival times; i.e., it converts the arrival times into i.i.d. uniform random variables.

The experiments in Kim and Whitt (2014b) showed that the KS test used by Brown et al. (2005) has much greater power against alternative processes with non-exponential interarrival-time distributions. Kim and Whitt (2014b) also found that Lewis (1965) had discovered a different data transformation due to Durbin (1961) to use after the CU transformation, and that the Lewis KS test consistently has more power than the log KS test from Brown et al. (2005) (although the difference is small compared to the improvement over the CU KS test). Evidently the Lewis test is effective because it brings the focus back to the interarrival times. Indeed, the first step of the Durbin (1961) transformation is to re-order the interarrival times of the uniform random variables in ascending order. We display the full transformation in Kim and Whitt (2013b).

Kim and Whitt (2014b) also found that the CU KS test of a PP should not be dismissed out of hand. Even though the CU KS test of a PP has remarkably little power against alternative processes with non-exponential interarrival-time distributions, the simulation experiments show that the CU KS test of a PP turns out to be relatively more effective against alternatives with dependent exponential interarrival times. The data transformations evidently make the other methods less effective in detecting dependence, because the re-ordering of the interarrival times weakens the dependence. Hence, here we concentrate on the Lewis and CU KS tests. For applications, we recommend applying both of these KS tests.

3.1.4 Remaining Issues in Applications

Unfortunately, it does not suffice to simply perform these KS tests to arrival data, because there are other complications with the data. Indeed, when we first applied the Lewis KS test to call center and hospital arrival data, we found that the Lewis KS test inappropriately rejected the NHPP property. In this chapter we address three further problems associated with applying the CU KS test and the Lewis refinement from Kim and Whitt (2014b) to service system arrival data. After applying these additional steps, we conclude that the arrival data we looked at are consistent with the NHPP property, but we would not draw any blanket conclusions. We think that it is appropriate to conduct statistical tests in each setting. Our analysis shows that this should be done with care.

First, we might inappropriately reject the NHPP hypothesis because of *data rounding*. Our experience indicates that it is common for arrival data to be rounded, e.g., to the nearest second. This often produces many 0-length interarrival times, which do not occur in an NHPP, and thus cause the Lewis KS test to reject the PP hypothesis. As in Brown et al. (2005), we find that inappropriate rejection can be avoided by un-rounding, which involves adding i.i.d. small uniform random variables to the rounded data. In §3.2 we conduct simulation experiments showing that rounding a PP leads to rejecting the PP hypothesis, and that un-rounding avoids it. We also conduct

experiments to verify that un-rounding does not change a non-PP into a PP, provided that the rounding is not too coarse. If the KS test rejects the PP hypothesis before the rounding and un-rounding, and if the rounding is not too coarse, then we conclude that the same will be true after the rounding and un-rounding.

Second, we might inappropriately reject the NHPP hypothesis because we *use inappropriate subintervals* over which the arrival rate function is to be regarded as approximately constant. In §3.3, we study how to choose these subintervals. As a first step, we make the assumption that the arrival-rate function can be reasonably well approximated by a piecewise-linear function. In service systems, non-constant linear arrival rates are often realistic because they can capture a rapidly rising arrival rate at the beginning of the day and a sharply decreasing arrival rate at the end of the day, as we illustrate in our call center examples. (It is important to note that some fundamental smoothness in the arrival rate function is being assumed; see §3.3.7 for more discussion.) Indeed, ways to fit linear arrival rate functions have been studied in Massey et al. (1996). However, we do not make use of this estimated arrival rate function in our final statistical test; we use it only as a means to construct an appropriate PC rate function to use in the KS test. We develop simple practical guidelines for selecting the subintervals.

Third, we might inappropriately reject the NHPP hypothesis because, in an effort to obtain a larger sample size, we might *inappropriately combine data from multiple days*. We might avoid the time-of-day effect and the day-of-the-week effect by collecting data from multiple weeks, but only from the same time of day on the same day of the week. Nevertheless, as discussed in §3.1.2, the arrival rate may vary substantially over these time intervals over multiple weeks. We may fail to recognize that, even though we look at the same time of day and the same day of the week, that data from multiple weeks may in fact have variable arrival rate. That is, there may be over-dispersion in the arrival data. It is often not difficult to test for such over-dispersion, using standard methods, provided that we remember to do so. Even better is to use more elaborate methods, as in Ibrahim et al. (2012) and references therein. If these tests do indeed find that there is such over-dispersion,

then we should not simply reject the NHPP hypothesis. Instead, the data may be consistent with i.i.d. samples of a Poisson process, but one for which the rate function should be regarded as random over different days (and thus a stochastic process). In §4 we discuss (mostly review) how to test for over-dispersion in the arrival data.

After investigating those three causes for inappropriately rejecting the NHPP hypothesis, in §3.5 and §3.6 we illustrate these methods with call center and hospital emergency department arrival data. We draw conclusions in §3.7. A a short online supplement (Kim and Whitt 2013b) and a longer appendix (Kim and Whitt 2013a) can be found online as well.

3.2 Data Rounding

A common feature of arrival data is that arrival times are rounded to the nearest second or even the nearest minute. For example, a customer who arrives at 11:15:25.04 and another customer who arrives at 11:15:25.55 may both be given the same arrival time stamp of 11:15:25 (rounding to seconds). That produces batch arrivals or, equivalently, interarrival times of length 0, which do not occur in an NHPP. If we do not take account of this feature, the KS test may inappropriately reject the NHPP null hypothesis.

The rounding problem can be addressed by having accurate arrival data without rounding, but often that is not possible, e.g., because the rounding is done in the measurement process. Nevertheless, as observed by Brown et al. (2005), it is not difficult to address the rounding problem in a reasonable practical way by appropriately un-rounding the rounded data. If the data has been rounded by truncating, then we can un-round by adding a random value to each observation. If the rounding truncated the fractional component of a second, then we add a random value uniformly distributed on the interval $[0, 1]$ seconds. We let these random values be i.i.d. It usually is straightforward to check if rounding has been done, and we would only un-round to un-do the rounding that we see.

3.2.1 The Need for Un-Rouding

To study rounding and un-rounding, we conducted simulation experiments. We first simulated 1000 replications of an NHPP with constant rate $\lambda = 1000$ (an ordinary PP) on the interval $[0, 6]$, with time measured in hours, so that a mean interarrival time is 3.6 seconds. We then apply the CU KS test and the Lewis KS test, as described in Kim and Whitt (2014b), to three versions of the simulated arrival data: (i) raw; as they are, (2) rounded; rounded to the nearest second, and (3) un-rounded; in which we first round to the nearest second and then afterwards un-round by adding uniform random variables on $[0, 1]$ divided by 3600 (since the units are hours and the rounding is to the nearest second) to the arrival times from (2), as was suggested in Brown et al. (2005).

Table 3.1 summarizes the results of the 1000 experiments. For each of the three forms of the data (raw, rounded and un-rounded) and two KS tests (CU and Lewis), we display the number of the 1000 KS tests that fail to reject the PP hypothesis at significance level $\alpha = 0.05$ (#P), the average p-values (ave[p-value]) and the average percentage of 0 values (ave[% 0]). The Lewis test consistently rejects the PP null hypothesis when the arrival data are rounded, but the CU KS test fails to do so. In fact, it is clearly appropriate to reject the PP null hypothesis when the data are rounded, because the rounding produces too many 0-valued interarrival times. Table 3.1 shows that the rounding turns 12.7% of the interarrival times into 0.

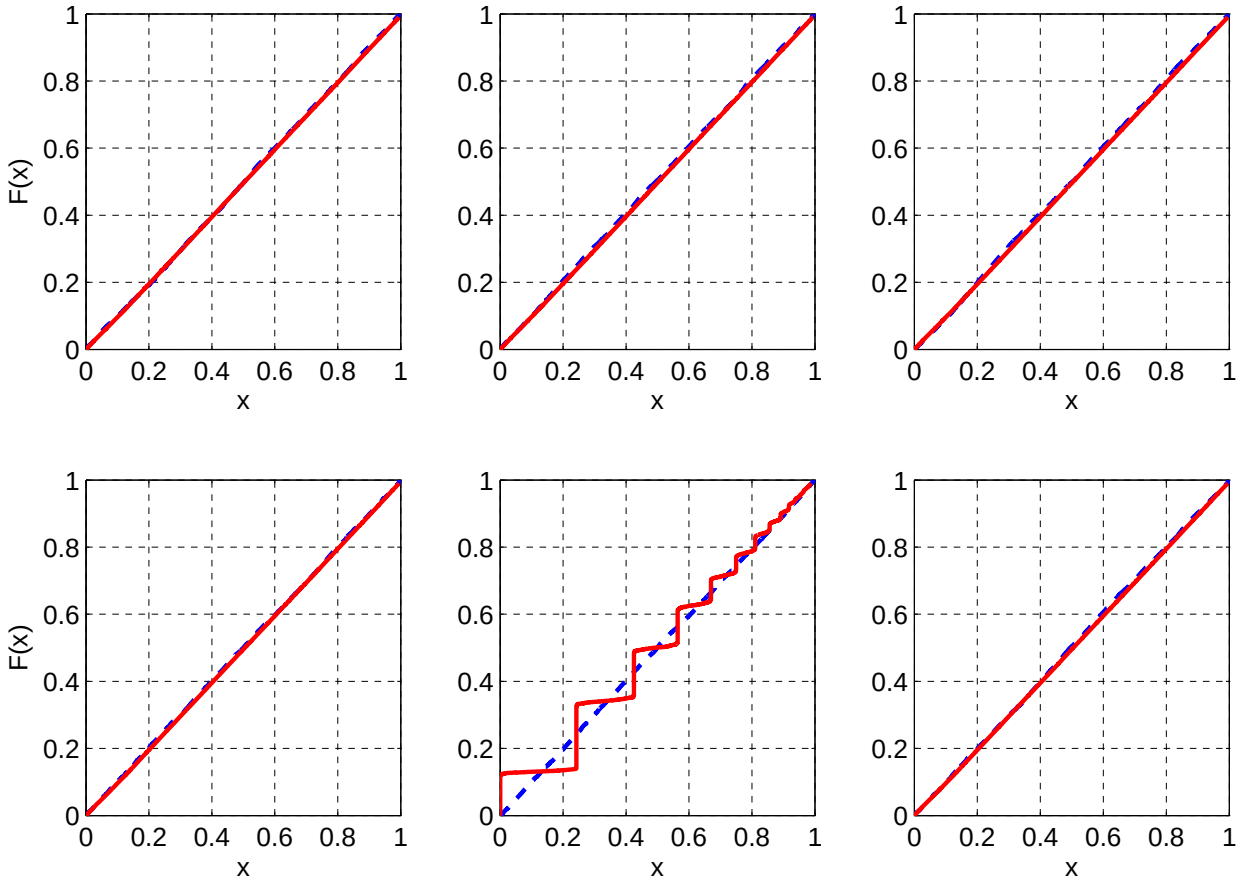
These test results illustrate the advantage of the Lewis KS test over the CU KS test. Since the Lewis test looks at the ordered interarrival times, all these 0-valued interarrival times are grouped together at the left end of the interval. As a consequence, the Lewis test strongly rejects the Poisson hypothesis when the data are rounded. Since the CU test looks at the data in order of the initial arrival times, the 0 interarrival times are spread out throughout the data and are not detected by the CU KS test. Fortunately, the problem of data rounding is well addressed by un-rounding. After the rounding, the Lewis KS test of a PP fails to reject the Poisson hypothesis when applied to a PP.

As in Kim and Whitt (2014b), we find that plots of the empirical cdf's used in the KS tests

Table 3.1: Results of the two KS tests with rounding and un-rounding: Poisson data

	CU			Lewis		
Type	# P	ave[p-value]	ave[% 0]	# P	ave[p-value]	ave[% 0]
Raw	944	0.50	0.0	955	0.50	0.0
Rounded	945	0.50	0.0	0	0.00	12.7
Un-rounded	945	0.50	0.0	961	0.50	0.0

Figure 3.1: Comparison of the average ecdf for a rate-1000 Poisson process. From top to bottom: CU, Lewis test. From left to right: Raw, Rounded, Un-rounded.



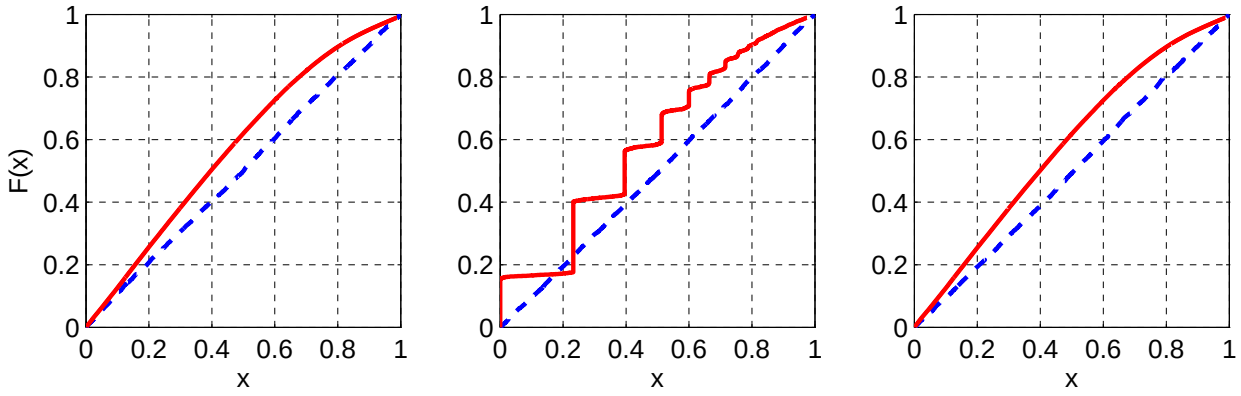
are very revealing. Figure 3.1 compares the average ecdf based on 100 replications of a rate-1000 Poisson process on the time interval [0,6] with the cdf of the null hypothesis. We do not show the 95% confidence intervals for the average ecdf's as they overlap with the average ecdf's. From

these plots, we clearly see that the Lewis test is very effective, whereas the CU KS test fails to detect any problem at all.

Table 3.2: Results of the two KS tests with rounding and un-rounding: $\mathbf{H_2}$ interarrival times

	CU			Lewis		
Type	# P	ave[p-value]	ave[% 0]	# P	ave[p-value]	ave[% 0]
Raw	705	0.21	0.0	0	0.00	0.0
Rounded	706	0.21	0.0	0	0.00	16.2
Un-rounded	706	0.21	0.0	0	0.00	0.0

Figure 3.2: Comparison of the average ecdf of a rate-1000 arrival process with H_2 interarrival times. Lewis test only. From left to right: Raw, Rounded, Un-rounded.



3.2.2 The Possible Loss of Power

Evidently, the rounding and subsequent un-rounding makes an arrival process more like an NHPP than it was before the rounding was performed. It is thus natural to wonder if the rounding and subsequent un-rounding causes a serious loss of power. To examine that issue, we performed additional experiments. First, we simulated 1000 replications of a renewal arrival process with constant rate $\lambda = 1000$ and i.i.d. hyperexponential (H_2; a mixture of two exponentials, and hence more variable than exponential) interarrival times X_j with the squared coefficient of variation

$c_X^2 = 2$ on the interval $[0, 6]$. The cdf of H_2 is $P(X \le x) \equiv 1 - p_1 e^{-\lambda_1 x} - p_2 e^{-\lambda_2 x}$. We further assume balanced means for ($p_1 \lambda_1^{-1} = p_2 \lambda_2^{-1}$) as in (3.7) of Whitt (1982) so that $p_i = [1 \pm \sqrt{(c_X^2 - 1)/(c_X^2 + 1)}]/2$ and $\lambda_i = 2p_i$. Table 1 of Kim and Whitt (2014b) shows that the Lewis test is usually able to detect this departure from the Poisson property and to reject the Poisson hypothesis.

Table 3.2 here shows the results of applying the CU test and the Lewis test to the renewal arrival process with H_2 interarrival times. We see that the Lewis KS test consistently rejects the Poisson hypothesis for the raw data, as it should, but the CU KS test fails to reject in 70% of the cases. Moreover, we observe that rounding and un-rounding does not eliminate the non-Poisson property. This non-Poisson property of the H_2 renewal process is detected by the Lewis test after the rounding and un-rounding. Figure 3.2 again provides dramatic visual support as well. (The ecdf's from the CU test look similar to the ones provided in Figure 3.1.)

We also conducted other experiments of the same kind to show that the un-rounding does not inappropriately cause the Lewis KS test to fail to reject a non-PP, provided that the un-rounding is not done too coarsely. Among the more interesting cases are two forms of batch Poisson processes. The first kind is a rate-1000 renewal process, in which the interarrival times are 0 with probability p and an exponential random variable with probability $1 - p$. The second is a modification of a PP in which every k^{th} arrival occurs in batches of size 2; the arrival rate is reduced to $1000k/(k + 1)$, so that the overall arrival rate is again 1000. Assuming that the rounding is done to the nearest seconds as in the PP and H_2 examples above, the un-rounding consistently detects the deviation from the PP when p is not too small (e.g., when $p \ge 0.05$) in the renewal process example and when k is not too large (e.g., when $k \le 9$) in the second modification of a PP with batches. As long as the rounding is not too coarse, the story for these examples is just like the H_2 renewal process we have already considered. The un-rounding removes all 0 interarrival times, but it still leaves too many very short interarrival times, so that the PP hypothesis is still rejected. The details appear in Kim and Whitt (2013b).

On the other hand, if the rounding is too coarse, as in the batch-Poisson examples above when the rounding is to the nearest minute instead of the nearest second, then the unrounding *can* hide the non-PP character of the original process, and thus reduce the power of the KS test. We also illustrate this phenomenon in Kim and Whitt (2013b). Overall, rounding should not matter, so that un-rounding is unnecessary, if the rounding is very fine, e.g., to less than 0.01 mean service time, while there is a danger of a loss of power if the rounding is too coarse, e.g., to more than a mean service time. We recommend using simulation to investigate in specific instances, as we have done here. See Kim and Whitt (2013b) and Kim and Whitt (2013a) for more discussion.

3.3 Choosing Subintervals With Nearly Constant Rate

In order to exploit the CU property to conduct KS tests of an NHPP, we assume that the rate function is approximately piecewise-constant (PC). Since the arrival rate evidently changes relatively slowly in applications, the PC assumption should be reasonable, provided that the subintervals are chosen appropriately. However, some care is needed, as we show in this section. Before starting, we should note that there are competing interests. Using shorter intervals makes the piecewise-constant approximation more likely to be valid, but interarrival times are necessarily truncated at boundary points and any dependence in the process from one interval to the next is lost when combining data from subintervals, so we would prefer longer subintervals unless the piecewise-constant approximation ceases to be appropriate.

As a reasonable practical first step, we propose approximating any given arrival rate function by a piecewise-linear arrival rate function with finitely many linear pieces. Ways to fit linear arrival rate functions were studied in Massey et al. (1996), and that can be extended to piecewise-linear arrival rate functions (e.g., by choosing roughly appropriate boundary times and applying the least-squares methods there over each subinterval with the endpoint values constrained). However, it usually should not be necessary to have a formal estimation procedure in order to obtain a suitable

rough approximation. In particular, we do not assume that we should necessarily consider the arrival rate function as fully known after this step; instead, we assume it is sufficiently well known to determine how to construct an appropriate PC approximation.

In this section we develop theory to support choosing subintervals for any given linear arrival rate function, which we *do* take as fully known. This theory leads to simple practical guidelines for evaluating whether (i) a constant approximation is appropriate for any given subinterval with linear rate and (ii) a piecewise-constant approximation is appropriate for any candidate partition of such a subinterval into further equally spaced subintervals; see §3.3.4 and §3.3.6, respectively. Equally spaced subintervals is only one choice, but the constant length is convenient to roughly judge the dependence among successive intervals.

3.3.1 A Call Center Example

We start by considering an example motivated by the banking call center data used in Kim and Whitt (2013e,f). For one 17-hour day, represented as $[6, 23]$ in hours, they produced the fitted arrival rate function

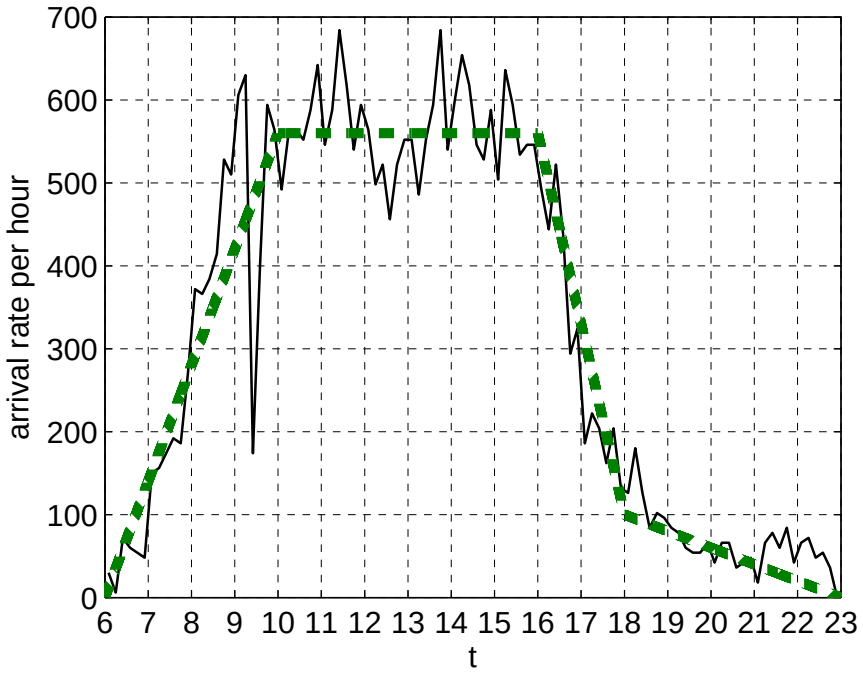
$$\lambda(t) = \begin{cases} 140(t-6) & \text{on [6,10],} \\ 560 & \text{on [10,16],} \\ 560 - 230(t-16) & \text{on [16,18],} \\ 100 - 20(t-18) & \text{on [18,23],} \end{cases} \tag{3.3}$$

as shown in Figure 3.3 (taken from Kim and Whitt (2013e,f)).

This fitted arrival rate function is actually constant in the subinterval $[10, 16]$, which of course presents no difficulty. However, as in many service systems, the arrival rate is increasing at the beginning of the day, as in the subinterval $[6, 10]$, and decreasing at the end of the day, as in the two intervals $[16, 18]$ and $[18, 23]$.

We start by considering an example motivated by Figure 3.3. The first interval $[6, 10]$ in Figure

Figure 3.3: Fitted piecewise-linear arrival rate function for the arrivals at a banking call center.



3.3 with linear increasing rate is evidently challenging. To capture the spirit of that case, we consider an NHPP with linear arrival rate function $\lambda(t) = 1000t/3$ on the interval $[0, 6]$. The expected total number of arrivals over this interval is 6000. We apply simulation to study what happens when we divide the interval $[0, 6]$ into $6/L$ equally spaced disjoint subintervals, each of length L, apply the CU construction to each subinterval separately and then afterwards combine all the data from the subintervals.

Table 3.3 and Figure 3.4 show the performance of the Lewis and CU KS tests as a function of the subinterval length. As before, #P is the number of KS tests passed at significance level $\alpha = 0.05$ out of 1000 replications. It shows the average p-values under ave[p-value] and the average percentage of 0 values in the transformed sequence under ave[$\%$ 0]. First, we see, just as in §3.2, that the Lewis test sees the rounding, but the CU test misses it completely. Second, we conclude that both KS tests will consistently detect this strong non-constant rate and *reject* the PP hypothesis with very high probability if we use $L = 6$ (the full interval $[0, 6]$) or even if $L = 1$ or

0.5. However, the Lewis KS tests will tend *not* to reject the PP hypothesis if we divide the interval into appropriately many equally spaced subintervals.

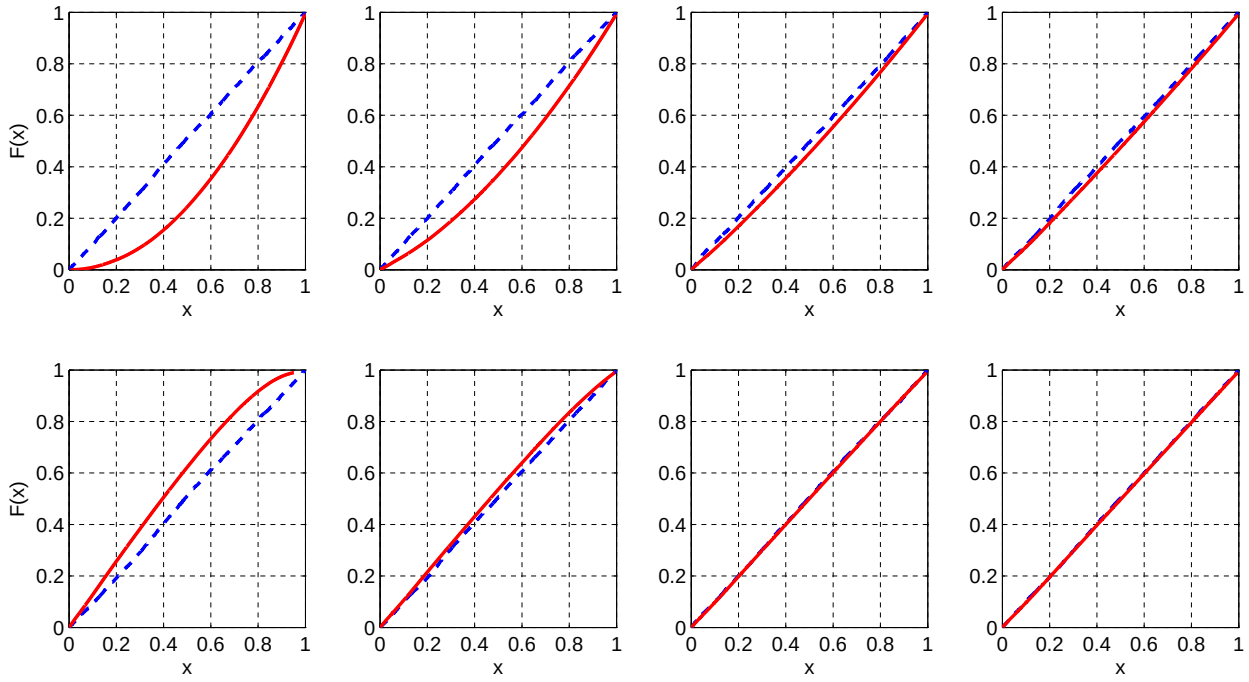
Since we are simulating an NHPP, the actual model differs from the PP null hypothesis only through time dependence. Consistent with the observations in Kim and Whitt (2014b), we see that the CU KS test is actually *more* effective in detecting this non-constant rate than the Lewis test. The non-constant rate produces a form of dependence, for which the CU test is relatively good. However, for our tests of the actual arrival data, we will wish to test departures from the NHPP assumption. Hence, we are primarily interested in the Lewis KS test. The results in §§3.3.3-3.3.6 below indicate that we could use $L = 0.5$ for the Lewis test.

Table 3.3: Performance of the alternative KS test of an NHPP as a function of the subinterval length **L**

		CU			Lewis		
L	Type	# P	ave[p-value]	ave[% 0]	# P	ave[p-value]	ave[% 0]
6	Raw	0	0.00	0.0	0	0.00	0.0
	Rounded	0	0.00	0.0	0	0.00	16.2
	Un-rounded	0	0.00	0.0	0	0.00	0.0
3	Raw	0	0.00	0.0	0	0.00	0.0
	Rounded	0	0.00	0.0	0	0.00	16.2
	Un-rounded	0	0.00	0.0	0	0.00	0.0
1	Raw	0	0.00	0.0	797	0.33	0.0
	Rounded	0	0.00	0.0	0	0.00	16.2
	Un-rounded	0	0.00	0.0	815	0.33	0.0
0.5	Raw	62	0.01	0.0	946	0.47	0.0
	Rounded	69	0.01	0.1	0	0.00	16.2
	Un-rounded	66	0.01	0.0	932	0.47	0.0
0.25	Raw	570	0.19	0.0	953	0.48	0.0
	Rounded	578	0.19	0.1	0	0.00	16.3
	Un-rounded	563	0.19	0.0	953	0.49	0.0

In the remainder of this section we develop theory that shows how to construct piecewise-

Figure 3.4: Comparison of the average ecdf of an NHPP with different subinterval lengths. From top to bottom: CU, Lewis test. From left to right: L = 6, 3, 1, 0.5.



constant approximations of the rate function. We then derive explicit formulas for the conditional cdf in three cases: (i) in general (which is complicated), (ii) when the arrival rate is linear (which is relatively simple) and (iii) when the data is obtained by combining data from equally spaced subintervals of a single interval with linear rate (which remains tractable). We then apply these results to determine when a piecewise-constant approximation can be considered appropriate for KS tests.

3.3.2 The Conditioning Property

We first observe that a generalization of the CU method applies to show that the scaled arrival times of a general NHPP, conditional on the number observed within any interval, can be regarded as i.i.d. random variables, but with a non-uniform cdf, which we call the *conditional cdf*, depending on the rate of the NHPP over that interval. That conditional cdf then becomes the asymptotic value of the conditional-uniform Kolmogorov-Smirnov test statistic applied to the arrival data as the

sample size increases, where the sample size increases by multiplying the arrival rate function by a constant.

Let $N \equiv \{N(t) : t \geq 0\}$ be an NHPP with arrival rate function λ over a time interval $[0, T]$. We assume that λ is integrable over the finite interval of interest and strictly positive except at finitely many points. Let Λ be the associated cumulative arrival rate function, defined by

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad 0 \leq t \leq T. \tag{3.4}$$

We will exploit a basic conditioning property of the NHPP, which follows by the same reasoning as for the homogeneous special case. It is significant that this conditioning property is independent of scale, i.e., it is unchanged if the arrival rate function λ is multiplied by a constant. We thus later consider asymptotics in which the sample size increases in that way.

Theorem 3.1 (*NHPP conditioning property*) *Let N be an NHPP with arrival rate function $c\lambda$, where c is an arbitrary positive constant. Conditional upon $N(T) = n$ for the NHPP N with arrival rate function $c\lambda$, the n ordered arrival times X_j, $1 \leq j \leq n$, when each is divided by the interval length T, are distributed as the order statistics associated with n i.i.d. random variables on the unit interval $[0, 1]$, each with cumulative distribution function (cdf) F and probability density function (pdf) f, where*

$$F(t) \equiv \Lambda(tT)/\Lambda(T) \quad \textit{and} \quad f(t) \equiv T\lambda(tT)/\Lambda(T), \quad 0 \leq t \leq 1. \tag{3.5}$$

In particular, the cdf F is independent of c.

We call the cdf F in (3.5) the *conditional cdf* associated with $N \equiv N(c\lambda, T)$. Let X_j be the j^{th} ordered arrival time in N over $[0, T]$, $1 \leq j \leq n$, assuming that we have observed $n \geq 1$ points in

the interval $[0, T]$. Let $\bar{F}_n(x)$ be the *empirical cdf* (ecdf) after scaling by dividing by T, defined by

$$\bar{F}_n(t) \equiv \frac{1}{n} \sum_{k=1}^{n} 1_{\{(X_j/T) \le t\}}, \quad 0 \le t \le 1. \tag{3.6}$$

We naturally are more likely to obtain larger and larger values of n if we increase the scaling constant c.

Observe that the ecdf $\{\bar{F}_n(t) : 0 \le t \le 1\}$ is a stochastic process with

$$E[\bar{F}_n(t)] = F(t) \quad \text{for all} \quad t, \quad 0 \le t \le 1, \tag{3.7}$$

where F is the conditional cdf in (3.5). As a consequence of Lemma 3.1 below and the Glivenko-Cantelli theorem, we immediately obtain the following asymptotic result.

Theorem 3.2 (*limit for empirical cdf*) *Assuming a NHPP with arrival rate function $c\lambda$, where c is a scaling constant, the empirical cdf of the scaled order statistics in (3.6), obtained after conditioning on observing n points in the interval $[0, T]$ and dividing by T, converges uniformly w.p.1 as $n \to \infty$ (which may be obtained from increasing the scaling constant c) to the conditional cdf F in (3.5), i.e.,*

$$\sup_{0 \le t \le 1} |\bar{F}_n(t) - F(t)| \to 0 \quad \textit{as} \quad n \to \infty. \tag{3.8}$$

We will usually omit the scaling constant in our discussion, but with the understanding that it always can be introduced. Since we want to see how the NHPP fares in a KS test of a PP, it is natural to measure the *degree of nonhomogeneity in the NHPP* by

$$\nu(NHPP) \equiv \nu(\lambda, T) = D \equiv \sup_{0 \le t \le 1} |F(t) - t|, \tag{3.9}$$

where F is the conditional cdf in (3.5). The degree of nonhomogeneity is closely related to the CU KS test statistic for the test of a PP, which is the absolute difference between the ecdf and the

uniform cdf, i.e.,

$$D_n \equiv \sup_{0 \le t \le 1} |\bar{F}_n(t) - t|; \tag{3.10}$$

see Marsaglia et al. (2003), Massey (1951), Miller (1956), Simard and L'Ecuyer (2011).

As a consequence of Theorem 3.2, we can describe the behavior of the conditional-uniform (CU) KS test of a Poisson process applied to a NHPP with general arrival rate function λ.

Theorem 3.3 (*limit of the KS test of a Poisson process applied to an NHPP*) *As* $n \to \infty$ *in a NHPP with rate function* λ *over* $[0, T]$,

$$D_n \to D \equiv \sup_{0 \le t \le 1} |F(t) - t|, \tag{3.11}$$

where D_n *is the CU KS test statistic in* (3.10) *and* D *is the degree of nonhomogeneity in* (3.9) *involving the the conditional cdf* F *in* (3.5).

Corollary 3.1 (*asymptotic rejection of the Poisson process hypothesis if NHPP is not a Poisson process*) *The probability that an NHPP with rate function* $n\lambda$ *will be rejected by the CU KS test for a PP converges to* 1 *as the scaling parameter* $n \to \infty$ *if and only if the* λ *is not constant w.p.1, i.e., if and only if the NHPP is not actually a PP.*

Proof. It is easy to see that the cdf F in (3.5) coincides with the uniform cdf t if and only if $\lambda(t)$ is constant. ∎

Corollary 3.1 suggests that a piecewise-constant approximation of a non-PP NHPP never makes sense with enough data, but we develop a positive result exploiting appropriate subintervals, where the number of subintervals grows with the sample size n; see Theorem 3.6.

3.3.3 An NHPP with Linear Arrival Rate Function

We now consider the special case of an NHPP with linear arrival rate function

$$\lambda(t) = a + bt, \quad 0 \le t \le T, \tag{3.12}$$

The analysis is essentially the same for increasing and decreasing arrival rate functions, so that we will assume that the arrival rate function is increasing, i.e., $b \ge 0$. There are two cases: $a > 0$ and $a = 0$; we shall consider them both. If $a > 0$, then cumulative arrival rate function can be expressed as

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds = at + \frac{bt^2}{2} = a\left(t + \frac{rt^2}{2}\right) \tag{3.13}$$

where $r \equiv b/a$ is the *relative slope*. If $a = 0$, then $\Lambda(t) = bt^2/2$.

Theorem 3.4 (*aymptotic maximum absolute difference in the linear case*) *Consider an NHPP with linear arrival rate function in* (3.12) *observed over the interval* $[0, T]$. *If* $a > 0$, *then the conditional cdf in* (3.5) *assumes the form*

$$F(t) = \frac{tT + (r(tT)^2/2)}{T + (rT^2/2)}, \quad 0 \le t \le 1; \tag{3.14}$$

if $a = 0$, *then*

$$F(t) = t^2, \quad 0 \le t \le 1. \tag{3.15}$$

Thus, if $a > 0$, *then the degree of nonhomogeneity of the NHPP can be expressed explicitly as*

$$D \equiv D(rT) \equiv \sup_{0 \le t \le 1} \{|F(t) - t|\} = |F(1/2) - 1/2| = \frac{1}{2} - \frac{\left(\frac{T}{2} + \frac{rT^2}{8}\right)}{\left(T + \frac{rT^2}{2}\right)} = \frac{rT}{8 + 4rT}. \tag{3.16}$$

If $a = 0$, *then* $D = 1/4$ (*which agrees with* (3.16) *when* $r = \infty$).

Proof. For (3.16), observe that $|F(t) - t|$ is maximized where $f(t) = 1$, so that it is maximized at $t = 1/2$. ∎

3.3.4 Practical Guidelines for a Single Interval

We can apply formula (3.16) in Theorem 3.4 to judge whether an NHPP with linear rate over an interval should be close enough to a PP with constant rate. (We see that should never be the case for a single interval with $a = 0$ because then $D = 1/4$.) In particular, the rate function can be regarded as approximately constant if the ratio $D/\delta(n, \alpha)$ is sufficiently small, where D is the degree of homogeneity in (3.16) and $\delta(n, \alpha)$ is the critical value of the KS test statistic D_n with sample size n and significance level α, which we always take to be $\alpha = 0.05$. Before looking at data, we can estimate n by the expected total number of arrivals over the interval.

We have conducted simulation experiments to determine when the ratio $D/\delta(n, \alpha)$ is sufficiently small that the KS test of a PP applied to an NHPP with that rate function consistently rejects the PP null hypothesis with probability approximately $\alpha = 0.05$. Our simulation experiments indicate that a ratio of 0.10 (0.50) should be sufficiently small for the CU (Lewis) KS test with a significance level of $\alpha = 0.05$.

Table 3.4 illustrates by showing the values of D, $\delta(n, \alpha)$ and $D/\delta(n, \alpha)$ along with the test results for selected subintervals of the initial example with $\lambda(t) = 1000t/3$ on the time interval $[0, 6]$. (The full table with all intervals and other examples appear in Kim and Whitt (2013a).)

3.3.5 Subintervals for an NHPP with Linear Arrival Rate

In this section we see the consequence of dividing the interval $[0, T]$ into k equal subintervals when the arrival rate function is linear over $[0, T]$ as in §3.3.3. As in the CU KS test discussed in Kim and Whitt (2014b), we treat each interval separately and combine all the data. An important initial observations is that the final cdf F can be expressed in terms of the cdf's F_j associated with the k

Table 3.4: Judging when the rate is approximately constant: the ratio $D/\delta(n,\alpha)$ for single subintervals with $\alpha = 0.05$

					ave		CU		Lewis	
L	Interval	$ave[n]$	r	D	$\delta(n,\alpha)$	$\frac{D}{ave[\delta(n,\alpha)]}$	# P	ave[p-val]	# P	ave[p-val]
6	[0,6]	5997	∞	0.250	0.018	14.28	0	0.00	0	0.00
3	[0,3]	1499	∞	0.250	0.035	7.15	0	0.00	0	0.00
	[3,6]	4499	0.33	0.083	0.020	4.12	0	0.00	481	0.15
1	[0,1]	167	∞	0.250	0.104	2.40	0	0.00	46	0.01
	[1,2]	450	1.00	0.083	0.060	1.38	22	0.01	896	0.43
	[2,3]	832	0.50	0.050	0.047	1.07	145	0.03	928	0.48
	[3,4]	1167	0.33	0.036	0.040	0.90	300	0.08	931	0.49
	[4,5]	1501	0.25	0.028	0.035	0.79	358	0.09	949	0.49
	[5,6]	1831	0.20	0.023	0.032	0.72	453	0.13	948	0.49
0.5	[0,0.5]	42	∞	0.250	0.207	1.21	46	0.01	562	0.18
	[0.5,1]	125	2.00	0.083	0.121	0.69	479	0.14	918	0.48
	[1.5,2]	292	0.67	0.036	0.079	0.45	766	0.29	945	0.50
	[2.5,3]	457	0.40	0.023	0.063	0.36	833	0.35	960	0.51
	[3.5,4]	623	0.29	0.017	0.054	0.31	865	0.38	938	0.51
	[4.5,5]	792	0.22	0.013	0.048	0.27	882	0.41	936	0.50
	[5.5,6]	957	0.18	0.011	0.044	0.25	893	0.42	951	0.50
0.25	[0,0.25]	10	∞	0.250	0.418	0.60	588	0.17	888	0.42
	[0.25,0.5]	32	4.00	0.083	0.239	0.35	841	0.37	946	0.49
	[0.5,0.75]	52	2.00	0.050	0.187	0.27	885	0.41	943	0.49
	[0.75,1]	73	1.33	0.036	0.157	0.23	907	0.44	947	0.50
	[1.75,2]	157	0.57	0.017	0.108	0.15	924	0.48	940	0.49
	[2.75,3]	239	0.36	0.011	0.087	0.12	931	0.48	956	0.50
	[3.75,4]	322	0.27	0.008	0.075	0.11	941	0.47	946	0.50
	[4.75,5]	407	0.21	0.006	0.067	0.10	937	0.48	953	0.50
	[5.75,6]	489	0.17	0.005	0.061	0.09	941	0.50	943	0.50

subintervals. In particular, we have the following lemma.

Lemma 3.1 (*combining data from equally spaced subintervals*) *If we start with a general arrival rate function and divide the interval* $[0,T]$ *into* k *subintervals of length* T/k*, then we obtain i.i.d. random variables with a conditional cdf that is a convex combination of the conditional cdf's for*

the individual intervals, i.e.,

$$\begin{aligned} F(t) &= \sum_{j=1}^{k} p_j F_j(t), \quad 0 \le t \le 1, \quad \textit{where} \\ F_j(t) &= \frac{\Lambda_j(tT/k)}{\Lambda_j(T/k)}, \quad 0 \le t \le 1, \quad 1 \le j \le k, \\ \Lambda_j(t) &= \Lambda(((j-1)T/k) + t) - \Lambda((j-1)T/k), \quad 0 \le t \le T/k, \quad 1 \le j \le k, \\ p_j &= \frac{\Lambda(jT/k) - \Lambda((j-1)T/k)}{\Lambda(T)}, \quad 1 \le j \le k. \end{aligned} \tag{3.17}$$

For the special case of a linear arrival rate function as in (3.12) *with* $a > 0$*,*

$$\begin{aligned} \Lambda_j(t) &= \frac{at(k(2+rt) + 2(j-1)rT)}{2k}, \quad 0 \le t \le T/k, \quad 1 \le j \le k, \\ F_j(t) &= \frac{t(2k + (2j-2+t)rT)}{2k + (2j-1)rT}, \quad 0 \le t \le 1, \quad 1 \le j \le k, \\ p_j &= \frac{2k + (2j-1)rT}{k^2(2+rT)}, \quad 1 \le j \le k, \quad \textit{and} \\ r_j &= \frac{b}{\lambda((j-1)T/k)} = \frac{bk}{a(k + (j-1)rT)}, \quad 1 \le j \le k. \end{aligned} \tag{3.18}$$

For the special case of a linear arrival rate function as in (3.12) *with* $a = 0$*,*

$$\begin{aligned} \Lambda_j(t) &= \frac{bt(kt + 2(j-1)T)}{2k}, \quad 0 \le t \le T/k, \quad 1 \le j \le k, \\ F_j(t) &= \frac{t(2j-2+t)}{2j-1}, \quad 0 \le t \le 1, \quad 1 \le j \le k, \\ p_j &= \frac{2j-1}{k^2}, \quad 1 \le j \le k, \quad \textit{and} \\ r_j &= \frac{k}{(j-1)T}, \quad 1 \le j \le k. \end{aligned} \tag{3.19}$$

We now apply Lemma 3.1 to obtain a simple characterization of the maximum difference from the uniform cdf when we combine the data from all the equally spaced subintervals.

Theorem 3.5 (*combining data from equally spaced subintervals*) *If we start with the linear arrival rate function in* (3.12) *and divide the interval* $[0, T]$ *into* k *subintervals of length* T/k*, and combine all the data, then we obtain*

$$D \equiv \sup_{0 \le t \le 1} \{|F(t) - t|\} = \sum_{j=1}^{k} p_j D_j = \sum_{j=1}^{k} p_j \sup_{0 \le t \le 1} \{|F_j(t) - t|\}. \tag{3.20}$$

If $a > 0$*, then*

$$D = \sum_{j=1}^{k} \frac{p_j r_j T/k}{8 + 4 r_j T/k} \le C/k \quad \textit{for all} \quad k \ge 1 \tag{3.21}$$

for a constant C*. If* $a = 0$*, then*

$$D = \frac{p_1}{4} + \sum_{j=2}^{k} \frac{p_j/(j-1)}{8 + 4/(j-1)} \le C/k \quad \textit{for all} \quad k \ge 1 \tag{3.22}$$

for a constant C*.*

Proof. By Theorem 3.4, by virtue of the linearity, for each $j \ge 1$, $|F_j(t) - t|$ is maximized at $t = 1/2$. Hence, the same is true for $|F(t) - t|$, where $F(t) = \sum_{j=1}^{k} p_j F_j(t)$, which gives us (3.20). For the final bound in (3.21), use $r_j \le 1 + (T/ka)$ for all j. For the final bound in (3.22), use $r_j = (j/(j-1))^2 \le 4$ for all $j \ge 2$ with $p_1 = 1/k^2$. ∎

3.3.6 Practical Guidelines for Dividing an Interval into Equal Subintervals

Paralleling §3.3.4, if the rate is strictly positive on the interval (or instead if it is 0 at one endpoint), then we can apply formula (3.21) (respectively, (3.22)) in Theorem 3.5 to judge whether the partition of a given interval with linear rate into equally spaced subintervals produces an appropriate PC approximation. As before, we look at the ratio $D/\delta(n, \alpha)$, requiring that it be less than 0.10 (0.50) for the CU (Lewis) KS test with significance level $\alpha = 0.05$, where now D is given by (3.21) or (3.22) and $\delta(n, \alpha)$ is again the critical value to the KS test, but now applied to all the

data, combining the data after the CU transformation is applied in each subinterval. In particular, n should be the total observed sample size or the total expected number of arrivals, adding over all subintervals.

We illustrate in Table 3.5 by showing the values of D and $D/\delta(n,\alpha)$ along with the test results for each subinterval of the initial example with $\lambda(t) = 1000t/3$ on the time interval $[0, 6]$, just as in Table 3.3. In all cases, $ave[n]$ is 5997.33, and hence the $ave[\delta(n,\alpha)]$ values are the same and are approximately 0.0175. (Again, more examples appear in Kim and Whitt (2013a).)

In summary, we present the following algorithm for choosing an appropriate subinterval length in order for a PC approximation of a linear arrival rate over some interval.

1. Given an interval T whose fitted arrival rate function is linear ($\lambda(t) = a + bt$), let n be the number of arrivals in that interval.

2. Compute the critical value of KS test, $\delta(n,\alpha)$. It can be approximated as $\delta(n,\alpha) \approx 1.36/\sqrt{n}$ if $n > 35$ and when we choose $\alpha = 0.05$ (see Simard and L'Ecuyer (2011) and references therein for other values of n and α).

3. Start with subinterval length $L = T/2$. Given L, compute *the degree of nonhomogeneity of the NHPP* D using (3.21) if $a > 0$ and using (3.22) if $a = 0$.

4. Compute $D/\delta(n,\alpha)$. Use bisection method to find the value of L that gives the ratio $D/\delta(n,\alpha)$ less than 0.10 (0.50) for the CU (Lewis) KS test.

3.3.7 Asymptotic Justification of Piecewise-Constant Approximation

We now present a limit theorem that provides useful insight into the performance of the CU KS test of a NHPP with linear rate. We start with a non-constant linear arrival rate function λ as in (3.12) and then scale it by multiplying it by n and letting $n \to \infty$. We show that as the scale

Table 3.5: Judging if a PC approximation is good for an interval divided into equal subintervals: the ratio $D/ave[\delta(n,\alpha)]$

			CU		Lewis	
L	D	$D/ave[\delta(n,\alpha)]$	# P	ave[p-val]	# P	ave[p-val]
6	0.2500	14.278	0	0.00	0	0.00
3	0.1250	7.139	0	0.00	0	0.00
1	0.0417	2.380	0	0.00	797	0.33
0.5	0.0208	1.190	62	0.01	946	0.47
0.25	0.0104	0.595	570	0.19	953	0.48
0.1	0.0042	0.238	896	0.43	955	0.48
0.09	0.0038	0.214	902	0.43	954	0.48
0.08	0.0033	0.190	914	0.45	948	0.48
0.07	0.0029	0.167	923	0.47	960	0.49
0.06	0.0025	0.143	927	0.47	941	0.49
0.05	0.0021	0.119	941	0.50	958	0.49
0.01	0.0004	0.024	953	0.50	948	0.48
0.005	0.0002	0.012	944	0.49	943	0.48
0.001	0.00004	0.002	952	0.50	959	0.49

increases, with the number of subintervals increasing as the scale increases appropriately, the KS test results will behave the same as if the NHPP had constant rate. We will reject if it should and fail to reject otherwise (with probability equal to the significance level). In particular, it suffices to use k_n equally spaced subintervals, where

$$\frac{k_n}{\sqrt{n}} \to \infty \quad \text{as} \quad n \to \infty. \tag{3.23}$$

In order to have the sample size in each interval also grow without bound, we also require that

$$\frac{n}{k_n} \to \infty \quad \text{as} \quad n \to \infty. \tag{3.24}$$

For example, $k_n = n^p$ satisfies both (3.23) and (3.24) if $1/2 < p < 1$.

Theorem 3.6 (*asymptotic justification of the piecewise-constant approximation of linear arrival rate functions*) *Suppose that we consider a non-constant linear arrival rate function over the fixed interval* $[0, T]$ *as above scaled by* n. *Suppose that we use the CU KS test with any specified significance level* α *based on combining data over* k_n *subintervals, each of width* T/k_n. *If conditions* (3.23) *and* (3.24) *hold, then the probability that the CU KS test of the hypothesis of a Poisson process will reject the NHPP converges to* α *as* $n \to \infty$. *On the other hand, if*

$$\frac{k_n}{\sqrt{n}} \to 0 \quad as \quad n \to \infty, \tag{3.25}$$

then the probability that the CU KS test of a Poisson process will reject the NHPP converges to 1 *as* $n \to \infty$.

Proof. Recall that the critical value $\delta(n, \alpha)$ of the CU KS test statistic D_n has the form $c_\alpha/\sqrt{n}$ as $n \to \infty$, where n is the sample size (see Simard and L'Ecuyer (2011)), and here the sample size is Kn for all n, where K is some constant. Let $D^{(n)}$ be D above as a function of the parameter n. Hence, we can compare the asymptotic behavior of $\delta(n, \alpha)$ to the asymptotic behavior of $D^{(n)}$, which has been determined above. Theorem 3.5 shows that $D^{(n)}$ is asymptotically of the form C/k_n. Hence, it suffices to compare k_n to $\sqrt{n}$ as in (3.23) and (3.25). ∎

In §5 of Kim and Whitt (2013b) we conduct a simulation experiment to illustrate Theorem 3.6. In §6 of Kim and Whitt (2013b) we also obtain an asymptotic result paralleling Theorem 3.6 for a piecewise-continuous arrival rate function where each piece is Lipschitz continuous.

3.4 Combining Data from Multiple Days: Possible Over-Dispersion

When the sample size is too small, it is natural to combine data from multiple days. For example, we may have hospital emergency department arrival data and we want to test whether the arrivals from 9am to 10am can be modeled as an NHPP. However, if there are only 10 arrivals in $[9, 10]$ on average, then data from one day alone will not be sufficient to test the PP property. A common way to address this problem is to combine data from multiple days; e.g., we can use all interarrival times in $[9, 10]$ from 20 weekdays, which will give us a sample size of about 200 interarrival times. From Kim and Whitt (2014b), we know that sample size is sufficient.

In call centers, as in many other service systems, it is well known that there typically is significant variation in the arrival rate over the hours of each day and even over different days of the week. It is thus common to estimate the arrival rate for each hour of the day and day of the week by looking at arrival data for specified hours and days of the week, using data from several successive weeks. The natural null hypothesis is that those counts over successive weeks are i.i.d. Poisson random variables. However, that null hypothesis should not be taken for granted. Indeed, experience indicates that there often is excessive variability over successive weeks. When that is found, we say that there is *over-dispersion* in the arrival data.

In some cases, over-dispersion can be explained by special holidays and/or seasonal trends in the arrival rate. The seasonal trends often can be identified by applying time-series methods. However, it can be the case that the observed over-dispersion is far greater than can be explained in those systematic ways, as we will illustrate for the arrival data from a banking call center in §3.5.

3.4.1 Directly Testing for Over-Dispersion

In the spirit of the rest of this chapter, we recommend directly testing whether or not there is over-dispersion in arrival data. The null hypothesis is that the hourly arrival counts at fixed hours on fixed days of week over a succession of weeks constitute independent Poisson random variables

with the same mean. A commonly used way to test if n observations $x_1, \ldots, x_n$ can be regarded as a sample from n i.i.d. Poisson random variables is the *dispersion test*, involving the statistic

$$\begin{aligned} \bar{D} &\equiv \bar{D}_n \equiv \frac{(n-1)\bar{\sigma}_n^2}{\bar{x}_n} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\bar{x}_n}, \quad \text{where} \\ \bar{\sigma}^2 &\equiv \bar{\sigma}_n^2 \equiv \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{n-1} \quad \text{and} \quad \bar{x} \equiv \bar{x}_n \equiv \frac{\sum_{i=1}^n x_i}{n}; \end{aligned} \tag{3.26}$$

e.g., see Kathirgamatamby (1953). Since we are concerned with excessive variability, we consider the one-sided test and reject if $\bar{D}_n > \delta(n, \alpha)$ where $P(\bar{D}_n > \delta(n, \alpha)|H_0) = \alpha$, again using $\alpha = 0.05$. Under the null hypothesis, $\bar{D}_n$ is distributed as χ^2_{n-1}, a chi-squared random variable with $n-1$ degrees of freedom, which in turn is distributed as the sum of squares of $n-1$ standard normal random variables. Thus, under the null hypothesis, $E[\bar{D}_n|H_0] = n-1$, $Var(\bar{D}_n|H_0) = 2(n-1)$ and $(\chi^2_n - n)/\sqrt{2n}$ converges to the standard normal as n increases. Thus $\delta(n, 0.05) = \chi^2_{n-1,0.95}$, the 95^{th} percentile of the χ^2_{n-1} distribution.

See Brown and hao (2002) for a discussion and comparison of several tests of the Poisson hypothesis. The dispersion test above is called the conditional chi-squared test in §3.3 there; it is shown to perform well along with a new test that they introduce, which is based on the statistic

$$\bar{D}^{bz} \equiv \bar{D}_n^{bz} \equiv 4 \sum_{i=1}^n (y_i - \bar{y}_n)^2 \quad \text{where} \quad y_i \equiv \sqrt{x_i + (3/8)}, \tag{3.27}$$

with $\bar{y}_n = n^{-1} \sum_{i=1}^n y_i$. Under the null hypothesis, $\bar{D}_n^{(bz)}$ is distributed as χ^2_{n-1} as well. We used both tests for over-dispersion and found that the results were very similar, so we only discuss $\bar{D}_n$ in (3.26); see the Kim and Whitt (2013a) for details.

3.4.2 Avoiding Over-Dispersion and Testing for it with KS Tests

An attractive feature of the KS tests based on the CU property is that we can avoid the over-dispersion problem while testing for an NHPP. We can avoid the over-dispersion problem by ap-

plying the CU property separately to intervals from different days and then afterwards combining all the data. When the CU property is applied in this way, the observations become i.i.d. uniform random variables, even if the rates of the NHPP's are different on different days, because the CU property is independent of the rate of each interval. Of course, when we apply KS tests based on the CU property in that way and conclude that the data is consistent with an NHPP, we have not yet ruled out different rates on different days, which might be modeled as a random arrival rate over any given day.

One way to test for such over-dispersion is to conduct the KS test based on the CU property by combining data from multiple days, by *both* (i) combining all the data before applying the CU property and (ii) applying the CU property to each day separately and then combining the data afterwards. If the data are consistent with an NHPP with fixed rate, then these two methods will give similar results. On the other hand, if there is significant over-dispersion, then the KS test will reject the NHPP hypothesis if all the data is combined before applying the CU property. By conducting both KS tests of a PP, we can distinguish among three alternatives: (i) PP with fixed rate, (ii) PP with random rate and (iii) neither of those.

3.5 Banking Call Center Arrival Data

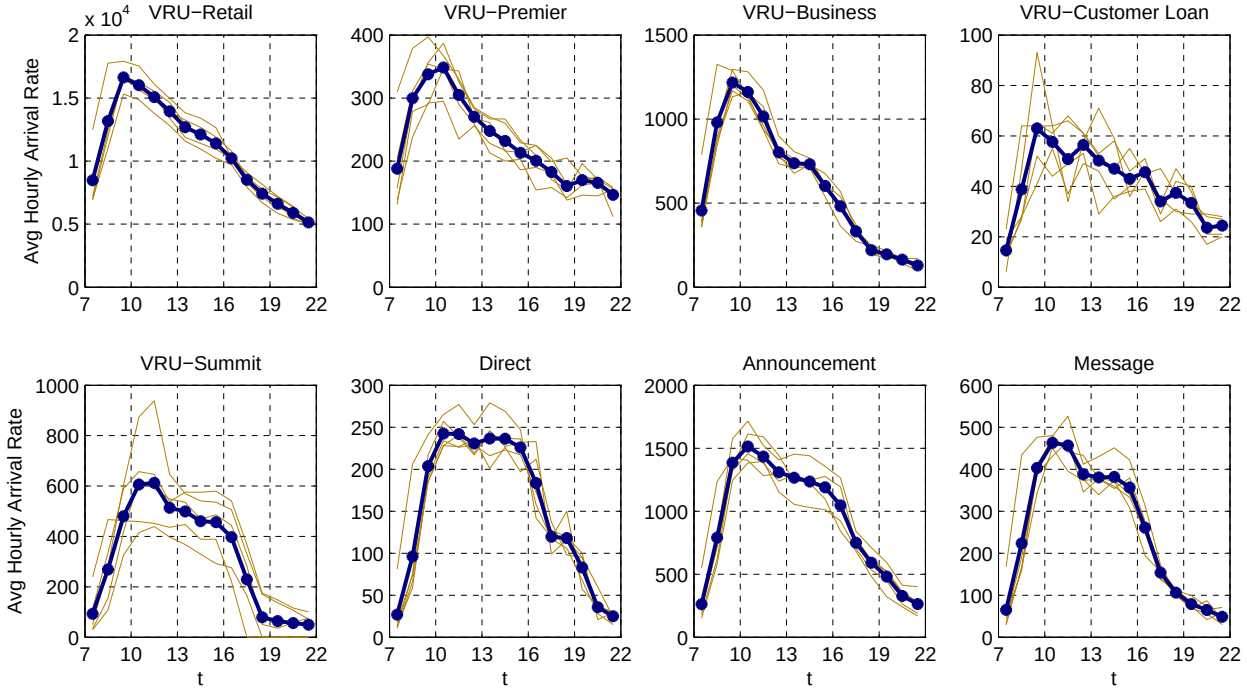
We now consider arrival data from service systems, first a banking call center and then, in the next section, a hospital emergency department. We use the same call center data used in Kim and Whitt (2013e,f), from a telephone call center of a medium-sized American bank from the data archive of Mandelbaum (2012), collected from March 26, 2001 to October 26, 2003. This banking call center had sites in New York, Pennsylvania, Rhode Island, and Massachusetts, which were integrated to form a single virtual call center. The virtual call center had 900 - 1200 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 Voice Response Unit (VRU) technology. In this study, we focus on arrival data during April 2001. There are 4 significant entry points to the system: through VRU ~92%, Announcement ~6%, Message ~1% and Direct group (callers that directly connect to an agent) ~1%; there are a very small number of outgoing and internal calls, and we are not including them. Furthermore, among the customers that arrive to the VRU, there are five customer types: Retail ~91.4%, Premier ~1.9%, Business ~4.4%, Customer Loan ~0.3%, and Summit ~2.0%.

3.5.1 Variation in the Arrival Rate Function

Figure 3.5 shows average hourly arrival rate as well as individual hourly arrival rate for each arrival type on Mondays.

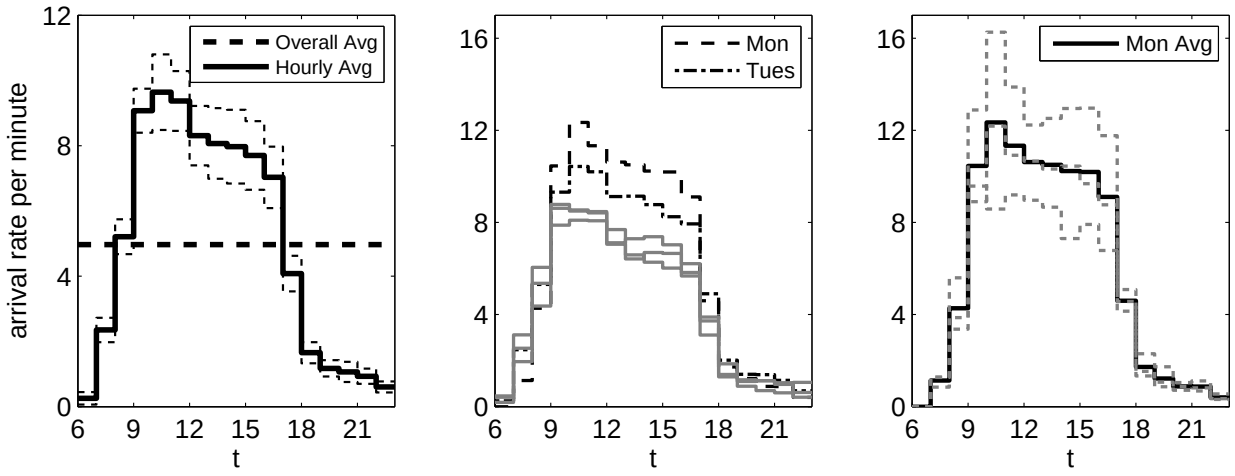
Figure 3.5: Call center arrivals: average and hourly arrival rates for 5 Mondays.



As usual, Figure 3.5 shows strong within-day variation. The variation over days of the week and over successive weeks in this call center data can be visualized by looking at four plots shown in Figure 3.6. The first plot on the left shows the average hourly arrival rate (per minute) over 18

weekdays (solid line) along with the daily average (horizontal dashed line). The average for each hour is the average of the arrival counts over that hour divided by 60 to get the average arrival rate per minute. The first plot also shows 95% confidence intervals about the hourly averages, which are quite wide in the middle of the day.

Figure 3.6: Average arrival rates. From left to right: Overall and hourly average arrival rates for 18 weekdays, average arrival rates by each day of week, arrival rates on three Mondays with its average.



Part of the variability seen in the first plot can be attributed to the day-of-the-week effect. That is shown by the second plot, which displays the hourly averages for the five weekdays. From this second plot, we see that the arrival rates on Mondays are the highest, followed by Tuesdays and then the others. Finally, the third plot focuses directly on the over-dispersion by displaying the hourly average rates by a specific day of the week, three Mondays. Even when restricting attention to a single day of the week, we see considerable variation.

However, it remains to quantify the variation. When we applied the dispersion test to the call center data, we found overwhelming evidence of over-dispersion. That is illustrated by the test results for 16 hours on 16 Fridays. Since the sample size for each hour was $n = 16$, $E[\bar{D}_n|H_0] = 15$ and $Var(\bar{D}_n|H_0) = 30$. The 95^{th} and 99^{th} percentiles of the χ^2_{15} distribution are, respectively, 25.0 and 30.6. However, the 16 observed values of $\bar{D}_n$ corresponding to the 16 hours on these Fridays

ranged from 163.4 to 1068.7, with an average of 356. The average value of $\bar{D}_n$ exceeds the 99^{th} percentile of the chi-squared distribution by a factor of 10. Moreover, the sample sizes were not small; the average hourly counts ranged from 29 to 503.

3.5.2 One Interval with Nearly Constant Arrival Rate

Figure 3.5 shows that the VRU - Summit arrival rate at the call center is nearly constant in the interval $[14, 15]$ (i.e., from 2pm to 3pm). We want to test whether the arrival process in $[14, 15]$ can be regarded as a PP. Consistent with the observations above, we see that there is a strong day-of-the-week effect. When we applied the dispersion test to all 30 days, we obtained $\bar{D}_n = 2320$, whereas $\delta(30, 0.05) = 43.8$. When we considered individual days of the week, we had 4 samples for each weekday and 5 for each weekend day. For Saturday we had $\bar{D}_n = 13.3$, while $\delta(5, 0.05) = 9.5$ and $\delta(5, 0.01) = 13.3$, showing that the p-value is 0.01, but in the other cases the $\bar{D}_n$ values ranged from 32.8 to 90.7, so that the arrival data for the time interval $[14, 15]$ on a fixed day of the week exhibits strong over-dispersion.

Since the arrival rate is approximately constant over $[14, 15]$, we do not need to consider subintervals in order to have a PC rate approximation. We can directly test for the PP, treating the data from separate days separately (and thus avoiding the day-of-the-week effect and the over-dispersion over successive weeks). First, we note that the arrival data were rounded to the nearest second. The results of the CU and Lewis KS tests, with and without un-rounding, are shown in Table 3.6. Table 3.6 shows that the Lewis test fails to reject the PP hypothesis in 29 out of 30 cases

Table 3.6: Results of KS Tests of PP for the interval $[14, 15]$

	Before un-rounding		Un-rounded	
Test	ave[p-val]	# P	ave[p-val]	# P
CU	0.54 ± 0.12	28	0.54 ± 0.12	28
Lewis	0.20 ± 0.08	19	0.49 ± 0.09	29

after un-rounding, but in only 19 before un-rounding. Just as in §3.2, the CU KS test fails to detect any problem caused by the rounding. Except for the over-dispersion, this analysis supports the PP hypothesis for the arrival data in the single interval $[14, 15]$.

3.5.3 One Interval with Increasing Arrival Rate

Figure 3.5 shows that the VRU - Summit arrival rate at the call center is nearly linear and increasing in the interval $[7, 10]$. We want to test whether the arrival process in $[7, 10]$ can be regarded as an NHPP.

Just as in the previous example, we see that there is a strong day-of-the-week effect. When we applied to dispersion test to all 30 days, we obtained $\bar{D}_n = 4257$, whereas $\delta(30, 0.05) = 43.8$. When we considered individual days of the week, we again had 4 samples for each weekday and 5 for each weekend day. For Wednesday we had $\bar{D}_n = 7.3$, while $\delta(4, 0.05) = 7.8$ and $\delta(4, 0.01) = 11.3$, and the p-value is 0.06, but for the other days of the week the $\bar{D}_n$ values ranged from 64.5 to 418, so that the arrival data for the time interval $[7, 10]$ on a fixed day of the week exhibits strong over-dispersion.

Since the arrival rate is nearly linear and increasing over $[7, 10]$, we need to use subintervals, as discussed in §3.3. Table 3.7 shows the result of using different subinterval lengths, L = 3, 1.5, 1, and 0.5 hours. The average number of arrivals over 30 days was 677.7 ± 111.1. We observe that more days pass the Lewis test as we decrease the subinterval lengths (and hence make the piecewise-constant approximation more appropriate in each subinterval). When we use L=0.5, all 30 days in April pass the Lewis test. We also see the importance of un-rounding; with L=0.5, only 18 days instead of 30 days pass the Lewis test when the arrival data are not un-rounded.

Table 3.7: Results of KS Tests of NHPP for the interval $[7, 10]$

		Before un-rounding		Un-rounded	
L (hours)	Test	ave[p-val]	# P	ave[p-val]	# P
3	CU	0.00 ± 0.00	0	0.00 ± 0.00	0
	Lewis	0.00 ± 0.01	1	0.04 ± 0.05	4
1.5	CU	0.02 ± 0.03	1	0.02 ± 0.03	1
	Lewis	0.09 ± 0.08	7	0.26 ± 0.11	18
1	CU	0.08 ± 0.04	12	0.08 ± 0.04	12
	Lewis	0.16 ± 0.08	15	0.48 ± 0.10	29
0.5	CU	0.23 ± 0.09	21	0.23 ± 0.10	21
	Lewis	0.20 ± 0.09	18	0.51 ± 0.10	30

3.5.4 The KS Test of All the Call Center Arrival Data

We now consider all the call center arrival data. Table 3.8 shows the result of applying the Lewis test to all the call center data by call type using subinterval length L equal to one hour to un-rounded arrival times (detailed results as well as CU test results can be found in Kim and Whitt (2013a)). We avoid the overdispersion by applying the CU transformation to all hours separately and then combining the data. The average number of observations, average p-value with associated 95% confidence intervals and the number of days (out of 30 days) that passed each test at significance level $\alpha = 0.05$ are shown.

The results of the tests lead us to conclude that the arrival data from all these groups of customers are consistent with the NHPP hypothesis, with the possible exception of the VRU-Retail group. We conjecture that the greater tendency to reject the NHPP hypothesis for the VRU-Retail group is due to its much larger sample size. To test that conjecture, we reduce the sample size. We do so by further dividing the time intervals into 3-hour long subintervals. Table 3.8 shows that we are much less likely to reject the NHPP null hypothesis when we do this.

In conclusion, we find significant over-dispersion in the call center data, i.e., variation over

Table 3.8: Lewis KS test applied to the call center data by type with $L = 1$ and un-rounding

	ave[# Obs]	ave[p-val]	# P
VRU-Retail	$1.4 \times 10^5 \pm 1.5 \times 10^4$	0.15 ± 0.09	11
VRU-Premier	$2.9 \times 10^3 \pm 2.6 \times 10^2$	0.49 ± 0.10	30
VRU-Business	$6.8 \times 10^3 \pm 1.2 \times 10^3$	0.49 ± 0.12	24
VRU-CL	$4.3 \times 10^2 \pm 5.7 \times 10^1$	0.44 ± 0.12	25
VRU-Summit	$3.3 \times 10^3 \pm 5.7 \times 10^2$	0.46 ± 0.10	28
Business	$1.5 \times 10^3 \pm 2.6 \times 10^2$	0.44 ± 0.12	25
Announcement	$9.7 \times 10^3 \pm 1.4 \times 10^3$	0.42 ± 0.13	22
Message	$2.6 \times 10^3 \pm 5.1 \times 10^2$	0.50 ± 0.11	30
VRU-Retail [7,10]	$3.5 \times 10^4 \pm 4.6 \times 10^3$	0.37 ± 0.12	22
VRU-Retail [10,13]	$3.7 \times 10^4 \pm 3.9 \times 10^3$	0.15 ± 0.08	13
VRU-Retail [13,16]	$2.8 \times 10^4 \pm 3.3 \times 10^3$	0.27 ± 0.11	20
VRU-Retail [16,19]	$2.1 \times 10^4 \pm 2.2 \times 10^3$	0.45 ± 0.11	27
VRU-Retail [19,22]	$1.4 \times 10^4 \pm 1.2 \times 10^3$	0.43 ± 0.11	27

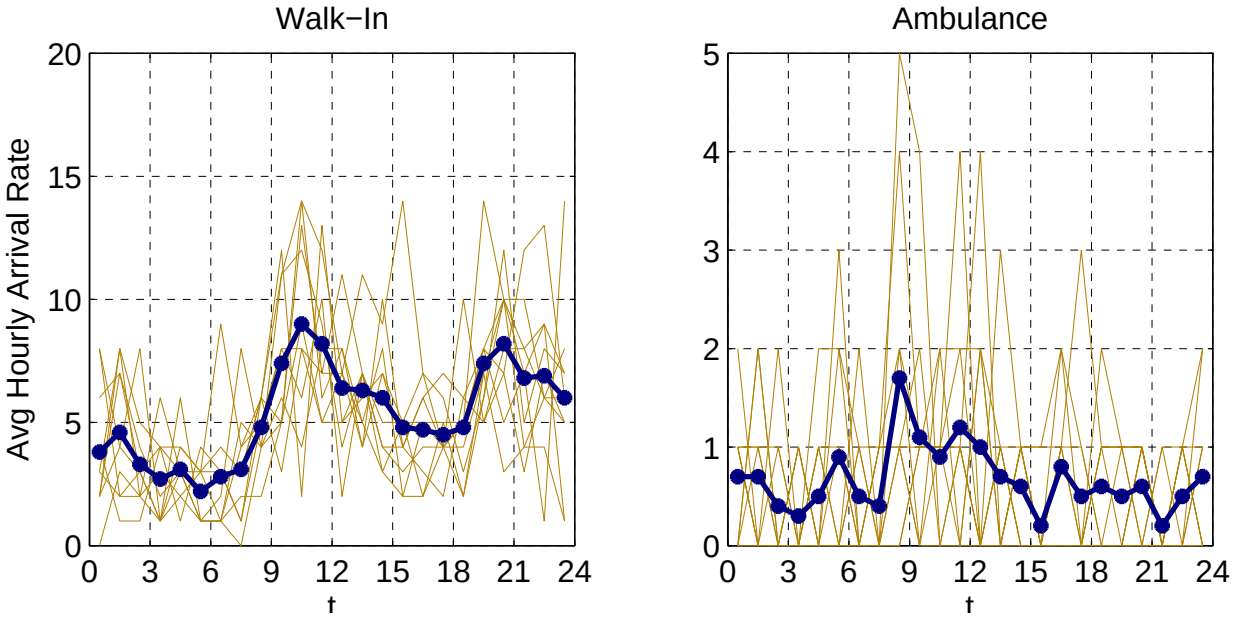
successive weeks in counts for the same hour on the same day of the week. Otherwise, we conclude that the arrival data is consistent with the NHPP hypothesis. However, failure to reject the NHPP null hypothesis depends critically on (i) un-rounding, (ii) properly choosing subintervals over which the rate can be regarded as approximately constant and (iii) avoiding the over-dispersion by applying the CU transformation to different hours separately.

3.6 Hospital Emergency Department Arrival Data

The emergency department (ED) arrival data are from one of the major teaching hospitals in South Korea, collected from September 1, 2012 to November 15, 2012. We focus on 70 days, from September 1, 2012 to November 9, 2012. There are two major entry groups, walk-ins and ambulance arrivals. On average, there are 138.5 arrivals each day with ~88% walk-ins and ~12% ambulance arrivals. Figure 3.7 shows the average hourly arrival rates for each arrival type on ten

Mondays.

Figure 3.7: Hospital ED arrivals: average and hourly arrival rates for 10 Mondays.



We observe less within-day variation among the ED arrivals than among the call center arrivals.

We first apply the dispersion test to test the Poisson hypothesis for daily counts for all days ($n = 70$) and all weekdays ($n = 50$), for all arrivals and by the two types. We can compare the dispersion statistic $\bar{D}$ values to $\chi^2_{n-1,1-\alpha}$ values: for each (n, α) pair: (70,0.05): 89.4, and (50,0.05): 66.3. The dispersion test rejects the Poisson hypothesis for the walk-in arrivals and the daily totals, with $448 \leq \bar{D}_n \leq 520$ in the 4 cases, while it does not reject for the ambulance arrivals, with $\bar{D}_n = 79.8$ for $n = 70$ (p-value 0.17) and $\bar{D}_n = 50.5$ for $n = 50$ (p-value 0.42).

However, in the analysis above we have not yet considered the day-of-the-week effect. When we analyze the walk-in arrivals by day of the week, we obtain $n = 10$ and $\chi^2_{n-1,1-\alpha} = 16.9$ for $(n, \alpha) = (10, 0.05)$. The observed daily values of $\bar{D}_n$ on the 7 days of the week, starting with Sunday, were 14.9, 9.4, 15.8, 10.7, 3.0, 25.3 and 14.4. Hence, we would reject the Poisson hypothesis only on the single day Friday. The associated p-values were 0.09, 0.40, 0.07, 0.30, 0.97, 0.00 and 0.11. While we might want to examine Fridays more closely, we tentatively conclude that there is no over-dispersion in the ED arrival data. We do not reject the Poisson hypothesis for

ambulance arrivals on all days and walk-in arrivals by day of the week.

Table 3.9: KS tests of NHPP for the hospital ED data

			Before un-rounding				Un-rounded			
			$L = 24$		$L = 1$		$L = 24$		$L = 1$	
Type	Day	n	CU	$Lewis$	CU	$Lewis$	CU	$Lewis$	CU	$Lewis$
Walk-In	Mon	1599	0.00	0.00	0.34	0.00	0.00	0.00	0.97	0.62
	Tues	1278	0.00	0.00	0.32	0.00	0.00	0.00	0.92	0.63
	Wed	1085	0.00	0.00	0.15	0.00	0.00	0.04	0.13	0.94
	Thurs	1063	0.00	0.00	0.58	0.00	0.00	0.02	0.68	0.36
	Fri	1122	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.54
	Sat	968	0.00	0.00	0.10	0.00	0.00	0.00	0.07	0.95
	Sun	1298	0.00	0.00	0.26	0.00	0.00	0.00	0.90	0.93
	Ave	1201.9	0.00	0.00	0.25	0.00	0.00	0.01	0.53	0.71
	# P		0/7	0/7	6/7	0/7	0/7	0/7	6/7	7/7
Ambulance	Mon	160	0.94	0.00	0.16	0.00	0.94	0.34	0.34	0.24
	Tues	162	0.08	0.00	0.19	0.00	0.08	0.69	0.16	0.88
	Wed	152	0.01	0.00	0.85	0.00	0.01	0.93	0.95	0.32
	Thurs	171	0.00	0.00	0.61	0.00	0.00	0.22	0.50	0.34
	Fri	169	0.03	0.00	0.00	0.00	0.03	0.71	0.01	0.28
	Sat	139	0.07	0.00	0.78	0.00	0.07	0.34	0.69	0.75
	Sun	192	0.15	0.00	0.35	0.00	0.15	0.46	0.48	0.08
	Ave	163.6	0.18	0.00	0.42	0.00	0.18	0.53	0.45	0.41
	# P		4/7	0/7	6/7	0/7	4/7	7/7	6/7	7/7

Next, we apply the CU and Lewis KS tests of an NHPP to the ED arrival data. First, based on the dispersion test results, we combine the data over the 10 weeks for each type and day of the week. We consider two cases for L: $L = 24$ (the entire day) and $L = 1$ using single hours as subintervals. Table 3.9 shows that, with un-rounding and subintervals of length $L = 1$, the Lewis test never rejects the PP hypothesis, while the CU test rejects only once (Fridays). As before, using un-rounding and subintervals is critical to these conclusions.

3.7 Conclusions

We examined call center and hospital arrival data and found that they are consistent with the NHPP hypothesis, i.e., that the KS tests of an NHPP applied to the data fail to reject that hypothesis, except that significant over-dispersion was found in the call center data. In particular: (i) variation in the arrival rate over the hours of each day was very strong for the call center data and significant for the ED data, (ii) variation in the arrival rate over the days of the week was significant for both the call center and ED data, except for ambulance arrivals, and (iii) variation in the arrival rate over successive weeks for the same time of day and day of week (over-dispersion) was significant for the call center data but not the ED data.

The analysis was not entirely straightforward. The majority of the chapter was devoted to three issues that need to be addressed and showing how to do so. §3.2 discussed data rounding, showing that its impact can be successfully removed by un-rounding. Consistent with Kim and Whitt (2014b), the Lewis test is highly sensitive to the rounding, while the CU KS test is not. §3.3 discussed the problem of choosing subintervals so that the PC rate function approximation is justified. Simple practical guidelines were given for (i) evaluating any given subinterval in §3.3.4 and (ii) choosing an appropriate number of equally spaced subintervals in §3.3.6. Again consistent with Kim and Whitt (2014b), the CU KS test is more sensitive to the deviation from a constant rate function than the Lewis KS test. Finally, §3.4 discussed the problem of over-dispersion caused by combining data from multiple days that do not have the same arrival rate. These three issues played an important role for both sets of data.

Chapter 4

Intensive Care Unit Admission Control

We examine the admission process to hospitals' Intensive Care Units (ICUs), which currently lacks well-defined admission criteria. A major challenge that has impeded the progress of developing ICU admission standards is that the impact of ICU admission on patient outcomes has not been well quantified, making it difficult to evaluate the performance of candidate admission strategies. Using a large patient-level dataset of over 190,000 hospitalizations across 15 hospitals, we first quantify the cost of denied ICU admission for a number of patient outcomes. We make methodological contributions in this context, improving upon previously developed instrumental variable approaches. Using the estimates from our econometric analysis, we provide a framework to evaluate the performance of various admission strategies. By simulating a hospital with 21 ICU beds, we then show that we could save about 1.9 million dollars per year by using our optimal objective policy designed to reduce readmissions and hospital length-of-stay. We also discuss the role of physicians' discretion on the performance of alternative admission strategies. This chapter is an edited version of a paper currently under revision, Kim et al. (2014).

4.1 Introduction

Intensive Care Units (ICUs) are specialized inpatient units that provide care for the most critically ill patients. They are extremely expensive to operate, consuming 15- 40% of hospital costs (Brilli et al. 2001, Halpern et al. 2007, Reis Miranda and Jegers 2012) despite comprising less than 10% of the inpatient beds in the U.S. (Joint Position Statement 1994, Halpern and Greenstein 1994). Most hospital ICUs operate near full capacity (Green 2003, Pronovost et al. 2004), making ICU beds a limited resource which must be rationed effectively. In this work, we examine what could be changed to improve the ICU admission decision process, how to generate the necessary information to help make these decisions, and how these decisions should vary under different scenarios.

The obvious criteria for ICU admission is that very sick and unstable patients should be treated in the ICU, while stable patients do not require ICU care. However, determining the most unstable patients is a complex task that is subject to high variability depending on the training and experience of the particular physician on staff (Mullan 2004, Boumendil et al. 2012, Chen et al. 2012). A critical care task force established ICU admission, discharge, and triage standards that are highly subjective in nature; the task force even admits that "[t]he criteria listed, while arrived at by consensus, are by necessity arbitrary" (Task Force of the American College of Critical Care Medicine, Society of Critical Care Medicine 1999). Indeed, the medical community has started to point to a need to develop systematic criteria for ICU care (Kaplan and Porter 2011, Chen et al. 2013), claiming that a primary reason for this gap is the general lack of objective criteria to characterize the benefit of different practices.

Our work takes an important step towards addressing this issue in that aims to estimate the benefit of ICU care for *all* medical patients admitted to the hospital through the Emergency Department (ED). We focus on patients admitted through the ED, who typically exhibits high uncertainty in the volume and severity of incoming patients, and whose care is the most likely to be affected by

not only each patient's medical severity but also hospital operational factors. For ethical reasons, it is not possible to run a field experiment to randomize ICU treatment to patients to estimate this benefit. Prior research has used observational data to measure the impact of ICU treatment on patient outcomes (e.g., Sprung et al. (1999), Shmueli et al. (2004), Simchen et al. (2004), Simpson et al. (2005), Iapichino et al. (2010), Kc and Terwiesch (2012), Louriz et al. (2012)). We also utilize data from 15 hospitals covering over 190,000 hospitalizations (of which we consider the admission decisions of over 70,000 patients).

Working with observational data to answer our questions brings an important econometric challenge: the decision to admit a patient to the ICU is endogenous and this can generate biases in estimating the benefit of ICU admission. Specifically, there are discretionary patient health severity factors which are accounted for by the deciding physicians but unobserved in the data; this unobservable information that goes into the admission decision will be positively correlated with ICU admission and adverse patient outcomes, generating a positive bias in the estimate of the causal effect of ICU care on patient outcomes. Kc and Terwiesch (2012) and Shmueli et al. (2004) propose using the congestion level of the ICU (which can affect patients' access to ICU care) as an instrumental variable (IV) to address this endogeneity problem. To be a valid IV, ICU congestion should affect patient outcomes only through its effect on the access to ICU care. But since hospital resources are shared among patients, a congested inpatient unit could directly impact the patient's recovery during his stay in the unit, invalidating the required exogeneity assumption of the IV. Unlike these prior studies, our data has detailed information on every unit each patient visits, which allows us to separate the effect of ICU congestion on the admission decision from its effect during the patient's hospitalization period, thereby validating the IV identification strategy. Based on these detailed data, we also construct and test additional IVs based on physician's behavioral aspects that influence the admission decision. Many U.S. hospitals have started to collect data similar to the one used in this work, and so the proposed methodology is applicable in other hospital settings. Our analysis shows that ICU care can reduce patient adverse health outcomes in the range of 30% to

75% (depending on the outcome). Moreover, the fact that our study covers 15 hospitals of different sizes, specialties, and locations helps to validate the robustness and generalizability of our results.

Equipped with these estimates, we compare the performance of various ICU admission strategies. We examine how the information available for decision making can impact performance. In particular, we examine the current admission criteria used by hospitals, which utilizes objective patient measures as well as doctors' discretionary information. This discretion has potential to be highly informative in assessing the costs of denying ICU admission but may be hard to record into objective patient metrics. As such we compare the performance of the current policy to an 'optimal policy' which is based on objective metrics alone. We use our estimated model of the hospital's current admission policy to simulate the current system and compare its relative performance vis-à-vis a system which uses our derived optimal policy. We find that the proposed optimal admission policy that uses objective patient severity metrics can outperform the current policy on certain measures of patient outcomes, but not all of them. For this reason, we also examine the benefit of the doctors' discretionary assessment of patient risk by examining an optimal policy which incorporates both objective and discretionary information. We find that in doing so, patient outcomes unilaterally improve, and we are able to capture the value of the doctors' discretionary information. Considering all the patient outcome measures that we study, a conservative estimate of the benefit of using the suggested policies at a single hospital translates into savings of patient bed hours on the order of 2.2 years, equivalent to US$1.9 million. This is approximately 5 times larger than the benefit that would be obtained by adding an additional bed to the ICU, excluding the costs of maintaining the extra bed.

In summary, we make the following key contributions:

- **Patient Outcomes:** In order to evaluate the performance of various admission policies, we require a quantification of the impact of ICU admission. Using a large patient-level dataset of over 190,000 hospitalizations across 15 hospitals, we quantify the cost of denied ICU admission for a number of patient outcomes including hospital LOS, hospital readmission, and

patient transfers to higher levels of care. We demonstrate that the impact of ICU admission is highly variable for different patients and different outcomes. Thus, it is important to have an understanding of all of these when making admission decisions. We also make methodological contributions in this context, improving upon previously developed instrumental variable approaches to address endogeneity biases that arise in this estimation problem.

- **Evaluation and Comparison of ICU admissions:** Based on the estimates from our econometric analysis, we are able to calibrate a simulation model, which we use to compare the performance of various admission strategies. We compare the derived optimal admission policies with the current hospital admission policies and find that in some circumstances it is useful to base admission decisions on objective metrics of patient risk alone where as in others the manner in which physicians incorporate objective and discretionary criteria in the admission decision can be beneficial. We are also able to quantify the benefit of discretionary information by examining how much patient outcomes improve when optimizing the admission decision based on both discretionary and objective criteria versus objective criteria alone.

The rest of this chapter is organized as follows. In Section 4.2, we provide a literature review. Section 4.3 describes the context of the problem and the data used in this empirical study. Section 4.4 develops the econometric model to estimate the effect of admission decisions on various patient outcomes. Section 4.5 provides our estimation results. Section 4.6 uses the empirical results to develop a simulation study to compare the performance under the current ICU admission policy used by hospitals with alternative approaches. Section 4.7 summarizes our main contributions and provides guidelines for future research.

4.2 Literature Review

There have been a number of works in healthcare Operations Management (OM) that study the effect of workload and congestion on healthcare productivity. On the empirical side, Kc and Terwiesch (2009) show that hospital congestion can accelerate patient transportation time within the hospital; Kuntz et al. (2014) examine the impact of hospital load on in-hospital mortality using the ides of safety tipping points; Green et al. (2013) find that nurse absenteeism rates in an ED are correlated with anticipated future nurse workload levels; Ramdas et al. (2012) and Kc and Staats (2012) study the impact of surgeon experience on outcomes; Jaeker and Tucker (2013) report that the length of inpatient stays depends on current workload as well as the predictability and the pressure level of the incoming workload; and Batt and Terwiesch (2012) find workload-dependent service times in the ED.

A more specific area of interest within this broader space is the study of mechanisms to manage ICU capacity. Several empirical studies have examined how hospitals utilize adaptive mechanisms to navigate periods of high ICU congestion. When a hospital does not have sufficient downstream bed capacity, surgical cases may be either delayed or canceled (Cady et al. 1995). When a new patient requires ICU care, but there is no available bed, he may be delayed and board in another unit, such as the ED or the post-anesthesia care unit (Ziser et al. 2002, Chalfin et al. 2007). An econometric study by Louriz et al. (2012) shows that a full ICU is the main factor associated with late ICU admission. Furthermore, Allon et al. (2013) shows that ED boarding caused by a congested ICU is an important factor driving ambulance diversion.

A mechanism that has received considerable attention from the OM and medical communities is to speed up the treatment of current ICU patients to accommodate new, potentially more critically ill patients. Anderson et al. (2011) investigate daily discharge rates from a surgical ICU at a large medical center, and find higher discharge rates on days with high utilization and more scheduled surgeries. Kc and Terwiesch (2012) study the effect of ICU occupancy level on discharge practices

in a cardiac surgical ICU. They find that congested ICUs tend to speed-up the treatment of their patients and that these affected patients tend to be readmitted to the ICU more frequently. We argue that admission and discharge decisions are fundamentally very different and that they utilize different information and criteria. Hence, the detailed understanding of the discharge decision established in Kc and Terwiesch (2012) cannot provide insight into the admission decision we study here.

Indeed, another alternative to manage ICU capacity that has been considered in the past and is considered here is to control the admission of patients. During periods of high congestion, some patients who may benefit from ICU care might be denied access because the ICU is full or all available beds are being reserved for more severe incoming patients. ICU congestion is an important factor affecting ICU admission decisions (Singer et al. 1983, Strauss et al. 1986, Vanhecke et al. 2008, Robert et al. 2012). Other studies have obtained similar results in international hospitals: Escher et al. (2004) in Switzerland, Azoulay et al. (2001) in France, Shmueli et al. (2004), Shmueli and Sprung (2005) and Simchen et al. (2004) in Israel, and Iapichino et al. (2010) in seven countries, including Italy, Canada, and the UK.

Most of the aforementioned studies on ICU admission control use patient severity measures which are based on scoring systems available only after patients are admitted to an ICU (Strand and Flaatten 2008). Examples include the Acute Physiology and Chronic Health Evaluation II (APACHE II) scores (Shmueli et al. 2004, Shmueli and Sprung 2005), Simplified Acute Physiology Score (SAPS II) (Iapichino et al. 2010, Simchen et al. 2004), Simplified Therapeutic Intervention Scoring System (TISS) (Simchen et al. 2004) and Mortality Prediction Model (MPM) (Louriz et al. 2012). These measures of patient severity are not available for a typical ED patient and hence, as argued by Franklin et al. (1990), they cannot be used to decide which patients should be routed to the ICU. In contrast, the hospitals we analyze use a uniform metric of patient severity available for all admitted patients: the Laboratory Acute Physiology Score (LAPS) (see Escobar et al. (2008) for details and validation of this metric). Previous work by Van Walraven et al. (2010) show that

LAPS is a reasonable predictor of patient length of stay and mortality. Utilizing this measure, we can analyze ICU admission decisions for all ED patients, and not just the patients who have been pre-screened for admission under subjective criteria, as done in prior work.

Closest to our work is Shmueli et al. (2003) that examines the impact of denied ICU admission on mortality. They consider patients who have already been referred for ICU admission and use an IV approach to measure how ICU admission decreases mortality for patients of different severity levels. Focusing on a sub-sample of patients pre-selected for ICU care has several drawbacks which we can address in our research design. **First**, Shmueli et al. (2003) use a severity measure (APACHE II) to measure the impact of ICU admission. This metric is generally assigned based on data available within the first 24 hours of ICU stay (Strand and Flaatten 2008), and so is not possible to use when considering which (of all) ED patients should be referred to the ICU. We instead develop admission criteria using metrics available to all patients in the ED. **Second**, their ICU admission criteria cannot be generalized to the (much larger) cohort of patients admitted from the ED. (In their study, 84% of patients are admitted to the ICU whereas in our sample, only 9.9% are admitted.) In particular, the benefit of ICU care may be exaggerated in Shmueli et al. (2003) because they only consider patients whose physicians have already determined that they require ICU care, whereas we are able to identify patients who will and will not benefit greatly from ICU care. **Third**, there is likely substantial variation in which patients will be recommended for ICU admission across hospitals and physicians due to heterogeneity in physicians' backgrounds, training, and opinions as documented in Mullan (2004), Weinstein et al. (2004), Fisher et al. (2004), O'Connor et al. (2004). In a sequel study to Shmueli et al. (2003), Shmueli and Sprung (2005) explicitly discuss that the admission policy in the ICU they are studying does not maximize the benefits of the ICU, and that "the discrepancies actually originate from [an] inappropriate referral policy." Our study provides criteria to use *before* any subjectivity in the pre-selection process can play a role. **Fourth**, we make important contributions by studying a number of different patient outcomes beyond mortality. This becomes important when the impact on mortality is similar

across many patients, but highly variable in other outcomes such as length-of-stay (LOS) and readmission. Accurately quantifying these effects is necessary when determining the optimal ICU admission decision.

We have seen that a number of mechanisms —including, but not limited to, ICU admission control— are used to manage ICU capacity in various settings. However, it is hard to find standards for when and how these mechanisms should be used; often there is substantial subjectivity in defining best practices. In a recent exploratory study, Chen et al. (2013) discuss the lack of standards in the field and point to a need to utilize Electronic Health Records to gain a better understanding of who benefits from ICU care in order to facilitate improved ICU triage decision making.

Indeed, our study utilizes data from a comprehensive Electronic Medical Records system. We focus on the ICU admission decision for patients that were admitted to the hospital through the ED to a medical service; in our data, about 55% (52%) of patient admitted to the hospital (ICU) are admitted via the ED to a medical service. The admission process works as follows. If an ED physician believes a patient is eligible for ICU admission, an intensivist will be called to the ED for consultation. While the intensivist has the ultimate decision about whether to admit the patient from the ED, the decision is typically a negotiation between the two physicians as to what the individual patient's needs are and what resources (e.g. ICU versus non-ICU beds) are available. The medical necessity of a patient plays a key role in the ICU admission decision, but the assessment of this necessity likely differs across physicians depending on his/her background and training (Mullan 2004, Weinstein et al. 2004, Fisher et al. 2004, O'Connor et al. 2004).

Our work takes an important step towards quantifying the benefits of ICU admission. Currently, most hospitals lack such measures, making it practically impossible to develop rigorous, evidence-based ICU care standards (Kaplan and Porter 2011). Although our focus is admission control, we conduct additional empirical analysis that accounts for other mechanisms mentioned above; see Section 4.5.2.

We draw upon the rich literature on the *stochastic knapsack problem* when considering optimal policies utilize our estimated quantified benefits of ICU admission: see Miller (1969), Weber and Stidham Jr (1987), Veatch and Wein (1992), Glasserman and Yao (1994), Papastavrou et al. (1996) and references there in. Specifically, we use a special case of the stochastic knapsack problem studied in Altman et al. (2001), and leverage some results from that work to characterize our optimal policy.

As we evaluate alternative admission policies in Section 4.6, we also discuss the role of physicians' discretion on the performance of alternative admission strategies. The value of discretionary criteria (or experts' input) in decision making has started receiving interest in other areas of Operations Management; e.g., see Anand and Mendelson (1997), Phillips et al. (2013) and Osadchiy et al. (2013). To the best of our knowledge, our study is the first to address this issue in the healthcare operations literature.

4.3 Setting and Data

We employ a large patient dataset collected from 15 hospitals, comprising of nearly 190,000 hospitalizations over the course of one and a half years. The hospitals are within an integrated healthcare delivery system, where insurers and providers fall under the same umbrella organization. The majority of patients treated within the system's hospitals are insured via this same organization. This allows us to ignore the potential impact insurance status may have on the care pathway of individual patients. However, we expect that our results can be extended to other hospitals that treat patients with heterogeneous insurance coverage.

In these 15 hospitals, inpatient units are broadly divided according to varying levels of nurse-to-patient ratios, treatment, and monitoring. The ICUs have a nurse-to-patient ratio of 1:1 to 1:2. There are two other kinds of inpatient units: general wards with ratios 1:3.5 to 1:4 and intermediate care units with ratios 1:2.5 to 1:3, though not all hospitals have intermediate care units. While

there is some differentiation within each level of care, the units are relatively fungible, so that if the medical ICU is very full, a patient may be admitted to the surgical ICU instead.

Patient-level information in our dataset includes patient age, gender, admitting diagnosis, hospital, two severity of illness scores–one based on lab results and comorbidities[1] and the other a predictor for in-hospital death [2]. In addition, we collect operational data that includes every unit each patient visits along with unit admission and discharge dates and times. Since we have an inpatient dataset, we do not have information on patients who are discharged directly from the ED.

Table 4.1: Description of the patient characteristics and seasonality control variables (labeled X_i in our econometric models) used to predict patient outcomes

Variable	Description and Coding
Age	Patient ages less than 39 were coded 1, 40-64 coded 2, 65-74 coded 3 (Medicare starts at 65), 75-84 coded 4 and above 85 coded 5
Gender	Females were coded 1 and males 0
Severity score 1: LAPS	Laboratory-based Acute Physiology Score (Escobar et al. 2008); measures physiologic derangement at admission and is mapped from 14 laboratory test results, such as arterial pH and white blood cell count, obtained in the 24 hours preceding hospitalization to an integer value that can range from 0 to a theoretical maximum of 256 (the maximum LAPS value in our data set was 166); coded as piecewise linear spline variables with knots at 39, 69, 89
Severity score 2: $\hat{P}$(Mortality)	an estimated probability of mortality (Escobar et al. 2008); predictors include LAPS and Comorbidity Point Score (measures the chronic illness burden and is based on 41 comorbidities); coded as piecewise linear spline variables with knots at 0.004, 0.075, 0.2
Admitting diagnosis	grouped into one of 44 broad diagnostic categories such as pneumonia; categorical variable to denote each diagnosis
Month/Time/Day	Month/Time/Day of week of ED admission; categorical variables

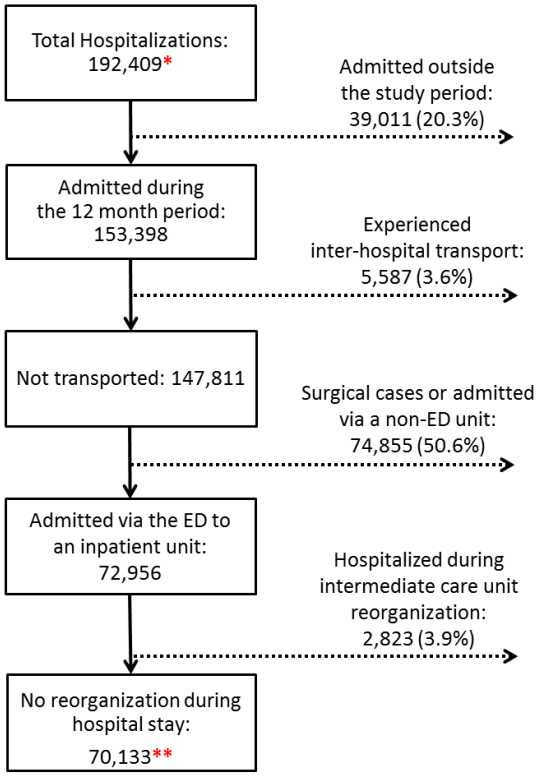


Figure 4.1: Selection of the patient sample

4.3.1 Data Selection

We now describe the sample selection procedure for the data used in this study as depicted in Figure 4.3.1. Hospitals in our dataset come from an integrated healthcare delivery system and had heterogeneous sizes of inpatient units. Because defining congestion in a small ICU is challenging and different mechanisms might be used to allocate beds in small ICUs, we consider only the patients who are treated in hospitals with ICUs of ten or more beds. There were 15 such hospitals; among them the maximum ICU occupancy varied from 10 to 44. The average percentage of ICU beds among inpatient beds was 12.9% with minimum of 9.3% and maximum of 21.5%.

We utilize patient flow data from all of 192,409 patient visits in the selected 15 hospitals–indicated by one star in Figure 4.3.1–to derive the capacity and instantaneous occupancy level

[1] i.e. chronic diseases, such as diabetes, that may complicate patient care and recovery.

[2] These multiple severity of illness scores reflect the complexity in defining objective severity of illness measures. Table 4.1 explains patient characteristics in detail.

of each inpatient unit. Because our dataset consists of patients admitted and discharged within the 1.5 year time period, we restrict our study to the 12 months in the center of the period to avoid censored estimation of capacity and occupancy. We exclude patients who experienced inter-hospital transport as it is difficult to determine whether it was due to medical or personal needs. Because of the reasons explained in Section 4.2, we focus on the patients who are admitted via the ED to a medical service. The sizes of the inpatient units were quite stable over our study period. However, four hospitals had a small change in the capacity of the intermediate care unit and we exclude patients who are hospitalized during these rare occurrences of intermediate care unit reorganization (such as reducing the number of beds). Our final dataset consists of 70,133 hospitalizations, as indicated by two stars in Figure 4.3.1.

4.3.2 Measuring Patient Outcomes

To quantify the benefit of ICU care, we focus on four types of patient outcomes whose summary statistics are provided in Table 4.3: (1) in-hospital death ($Mortality$), (2) hospital readmission ($Readmit$), (3) hospital length of stay (LOS) ($HospLOS$), and (4) transfer-up to a higher level of care ($TransferUp$). $Mortality$, $Readmit$, and $HospLOS$ are fairly standard patient outcomes used in the medical and OM communities (Iezzoni et al. 2003, Kc and Terwiesch 2009). We consider one additional measure of patient outcome, $TransferUp$, for the following reason. Typically, a patient will be transferred to an inpatient unit with lower level of care or be discharged from the hospital as his health state improves. Being transferred up to the ICU can be a sign of physiologic deterioration and such patients typically exhibit worse medical conditions (Luyt et al. 2007, Escobar et al. 2011). Accordingly, a $TransferUp$ event is defined as a patient's transfer to the ICU from an inpatient unit with lower level of care.[3] Note that patients who were admitted to and directly discharged from the ICU can never experience this event, and so we study

[3]ICU readmission, which qualifies as a $TransferUp$ event, has also been shown to lead to higher mortality and length of stay (Durbin Jr and Kopel 1993).

$TransferUp$ over the subset of patients who visited the general ward at least once during their hospital stay.

Defining readmission requires specifying a maximum elapsed time between consecutive hospital discharges and admissions. As this elapsed time increases, it becomes less likely that the complications were related to the care received during the initial hospitalization. Hence, after discussions with doctors, we define a relatively short time window for hospital readmission – within the first two weeks following hospital discharge. When analyzing $Readmit$, we did not include patients with in-hospital death as they cannot be readmitted.

We let $HospLOS$ measure the time from admission to the first inpatient unit until hospital discharge time, excluding the ED boarding time. A complication in analyzing $HospLOS$ is that its histogram reveals "spikes" every 24 hours. This is because of a narrow time-window for hospital discharge: more than 60% of the patients are discharged between 10am and 3pm, whereas admission times are less concentrated and demonstrate a markedly different distribution (a similar issue was reported in Armony et al. (2011) and Shi et al. (2012) on data from other hospitals). To avoid this source of measurement error, we measure $HospLOS$ as the number of nights the patient stayed in the hospital. In studying $HospLOS$, we include patients who died during their hospital stay. The results are similar if we exclude patients with in-hospital death.

4.4 Measuring the Impact of ICU Admission on Patient Outcomes

In this section, we study how access to ICU care affects patient mortality, readmissions, transfer-up events and hospital length of stay. Section 4.4.1 develops an econometric model to measure the impact of ICU care on these outcomes. The main challenge in this estimation is to account for the endogeneity in ICU admission decisions. Section 4.4.2 develops an estimation strategy using

Instrumental Variables (IVs) to address this endogeneity problem and Section 4.4.3 describes our final estimation models.

4.4.1 Econometric Model for Patient Outcomes

An ideal thought experiment to examine the implications of ICU admission on patient outcomes would be randomizing treatments to patients by allocating patients to the ICU and non-ICU units regardless of their severity condition. Of course, such an experiment would be impossible in practice due to ethical concerns. This limits us to work with observational data, which brings important challenges to the estimation, as we now describe.

Our unit of observation is a hospital visit of a patient, indexed by i. Let y_i denote a measure capturing a patient outcome of interest during this visit (e.g., $HospLOS_i$). There is extensive work in the medical literature that provides several patient severity measures that are useful in predicting patient outcomes. For example, Escobar et al. (2008) and Liu et al. (2010) illustrate how severity measures based on automated laboratory and comorbidity measures can be used to successfully predict in-hospital mortality and hospital length of stay. Let X_i denote those patient severity factors as well as seasonality controls that are observed in the data. We also control for hospitals, where $\omega_{h(i)}$ denotes the coefficients for a set of hospital indicator variables and $h(i)$ is patient i's hospital. Our main hypothesis is that ICU treatment has a causal effect on patient outcomes. Accordingly, let $Admit_i = 1$ if patient i is admitted to the ICU and zero otherwise. We model patient outcome y_i as a random variable with distribution $f(y_i|\beta_1, \beta_2, Admit_i, X_i, \omega_{h(i)})$, where the parameter β_1 captures the affect of ICU admission and β_2 measures the effect of the observable characteristics X_i on the patient outcome, respectively. For example, this distribution could be given by a model of the form:

$$\log(y_i) = \beta_1 Admit_i + X_i\beta_2 + \omega_{h(i)} + \varepsilon_i, \tag{4.1}$$

with the error term ε_i following a normal distribution so that y_i is log-normally distributed. In this

example, we have a linear regression with Gaussian errors, but our framework allows for more general specifications (e.g., binary patient outcomes).

The linear regression example (4.1) is useful to illustrate the main estimation challenge. A naive approach to estimate the effect of ICU admission on y_i is to estimate the regression model (4.1) via Ordinary Least Square (OLS) and interpret the estimate of β_1 as the causal effect of ICU admission on the outcome. This approach ignores that the admission decisions are endogenous; patient severity conditions that are unobservable in the data (e.g. the cognitive state of the patient) are likely to affect admission decisions. Figure 4.4 illustrates this endogeneity issue in further detail. The term ξ_i represents patient severity characteristics that are unobserved in the data but that are considered by the physicians when making the ICU admission decision. As such, both admission decisions and patient outcomes are affected by X_i and ξ_i. Since ξ_i is absorbed as part of the error term of model (4.1), the covariate $Admit_i$ is positively correlated with ε_i, therefore violating the strict exogeneity assumption required for consistent estimation through OLS. This endogeneity problem could introduce a positive bias in the estimate of the effect of ICU admission on patient outcomes, underestimating the value of ICU care (because we expect β_1 to be negative).

An alternative is to use Instrumental Variables (IVs) estimation to obtain consistent estimates of this linear regression model. A valid instrument should be correlated with the admission decision $Admit_i$ but unrelated to the unobserved patient severity factors ε_i determining the outcome y_i. We propose using hospital operational factors that affect the ICU admission decision but are otherwise unrelated to patient severity. We describe and validate these IVs in the next section.

4.4.2 Instrumental Variables

A valid instrumental variable, denoted by Z, needs to satisfy the following two conditions: (1) it has to influence the endogenous variable, in our case, the ICU admission decision $Admit_i$; and (2) it has to be exogenous, that is, it cannot affect the patient outcome measure y_i other than through

the admission decision. In this section, we discuss several potential instruments and empirically validates them here and in Section 4.5.

When deciding the ICU admission of an ED patient, the hospitals needs to evaluate the benefit of ICU treatment for this focal patient versus the opportunity cost of reserving the bed for a future, potentially more severe, incoming patient. This trade-off is particularly relevant when bed occupancy in the ICU is high – with only few beds left, admitting a patient now increases the probability that a future severe patient will be denied admission because the ICU is full. Because the number of beds is limited and the volume and severity of incoming patients is stochastic, the problem resembles an *admission control problem*: Altman et al. (2001) show that, for problems of this kind under various system conditions, the optimal admission control policy exhibits a reduction in the admission rate as the system occupancy increases.

We examine the data to identify differences in ICU admission rates due to occupancy. An ICU is labeled as "Busy" ($ICUBusy = 1$) if the bed occupancy is above the 95th percentile of its occupancy distribution, estimated by measuring the ICU bed occupancy every hour in the study period.[4] Figure 4.4.2 graphs the admission rates for 20 different patient groups (classified by their LAPS score on the horizontal axis) for two different occupancy levels: busy (marked with triangles) and not busy (marked with circles). The level of ICU occupancy associated to each patient is measured one hour prior to their ED discharge, which is a reasonable time period to cover the stage at which admission decisions are made. On top of the circles, we also show the percentage of the patients in each patient severity level group that saw an ICU that is "Not Busy". Across all groups, 90 to 92% patients are such patients, suggesting that there is no association between incoming patient severity level and the ICU occupancy level. Note that all 40 points in this graph have enough observations, with the smallest sample size being 144 patients. This figure shows that ICU admission decisions for patients at all severity levels are affected by ICU occupancy; among

[4]For instance, an ICU with 10 beds is considered busy when 9 or 10 of its beds are occupied if its occupancy is 8 beds or below 94.5% of the time and 9 beds or below more than 95% of the time.

patients in the same severity group, a lower percentage of patients who saw high ICU occupancy was sent to the ICU compared to the patients who saw low ICU occupancy level. We repeated the exercise for other cutoffs of ICU congestion including the 90^{th}, 85^{th} and 75^{th} percentiles. The change in admission rate was much smaller and non-existent for some groups of patients. Although other measures of ICU occupancy could be considered, Figure 4.4.2 suggests that $ICUBusy_i$ is a statistically powerful instrument, in the sense that it explains significant variation in the admission decision.

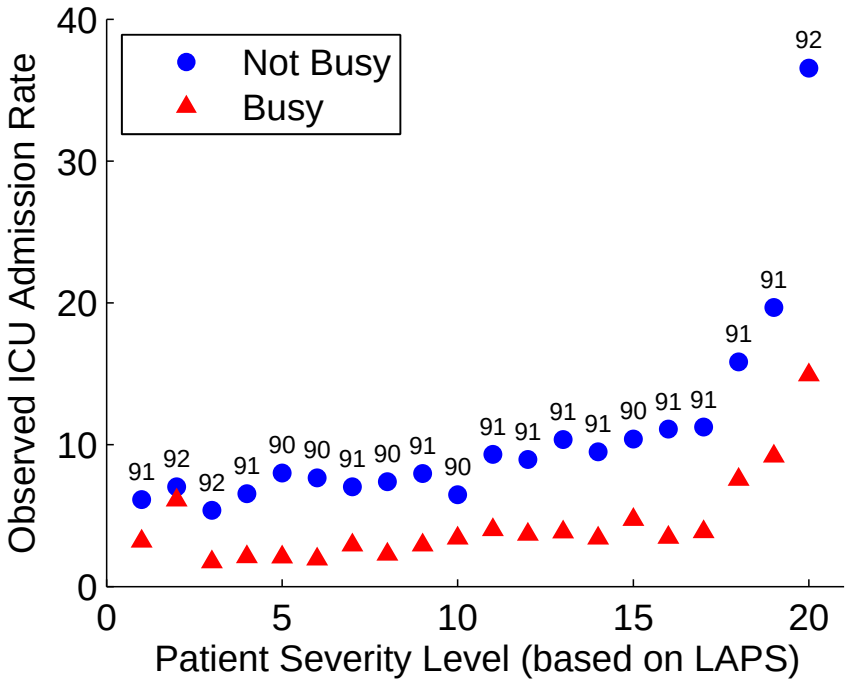


Figure 4.2: Observed ICU admission rate for patients with different severity levels characterized by LAPS, under high and low ICU occupancy (Busy and Not Busy, respectively). Numbers above circles indicate the fraction of patients (with given severity) that observed a "Not Busy ICU" one hour before their discharge from the ED.

However, for $ICUBusy_i$ to be a valid instrument it also has to be uncorrelated with the unobservable factors ε_i that affect patient outcomes. Kc and Terwiesch (2012) describe a potential mechanism that could lead to a violation of this assumption. They show that readmission rates tend to be higher for patients who experienced high ICU occupancy level during their ICU stay.

Moreover, the same effect could apply to other inpatient units visited by the patient.

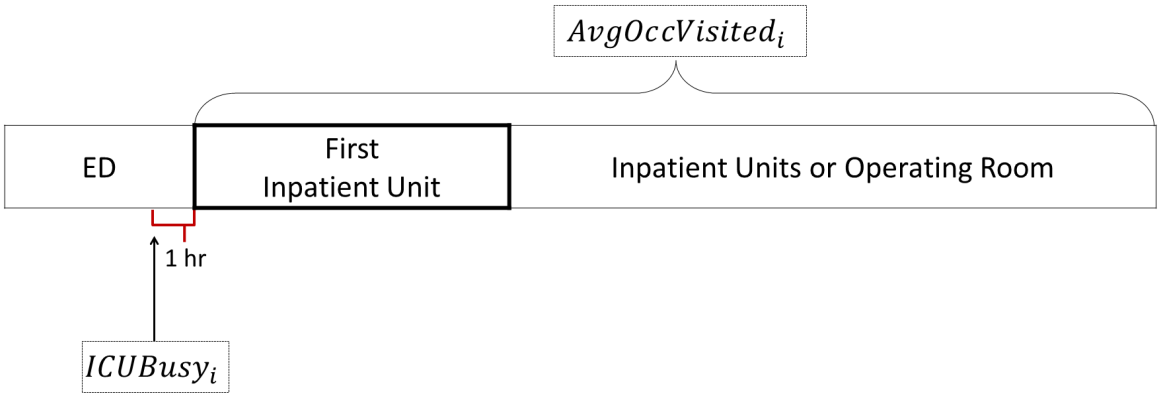


Figure 4.3: Time-Line of the process flow for patients admitted through the emergency department

To overcome this issue, we used the detailed information in our data about the complete care path of each patient to control for the congestion levels that a patient experienced in each of the visited inpatient units *during* his hospital stay. Specifically, let D_i be the set of days patient i stayed in the hospital (after leaving the ED) and $Occ_{i,d}$ the occupancy of the inpatient unit where patient i stayed in day d. The average occupancy of the inpatient units visited by the patient during his hospital stay is defined as $AvgOccVisited_i = \frac{1}{|D_i|}\sum_{d\in D_i} Occ_{i,d}$ (see Figure 4.3 for details on the time-line where this measure is calculated from).[5] We include $AvgOccVisited_i$ as an additional control variable in the outcome model (4.1) (in addition to the patient severity factors X_i). $AvgOccVisited_i$ is not perfectly correlated with $ICUBusy_i$ because the latter is measured *before* the patient is physically moved to the inpatient unit and the occupancy level typically varies during a patient's hospitalization period; the correlation between the two measures is 0.24 in our sample. Separating the effect of occupancy on admission decision from its effect

[5] We define capacity of an inpatient unit as the 95^{th} percentile of the bed occupancy distribution of that unit to compute $Occ_{i,d}$, because in many occasions, the maximal capacity is rarely observed as hospitals may temporary expand their standard capacity by a few beds in extreme circumstances (this was also pointed out in Armony et al. (2011) and Jaeker and Tucker (2013)). Given this definition, it is possible to have $Occ_{i,d}$ above 100%. The average $AvgOccVisited_i$ was 0.84 with median of 0.86 in our dataset

during the inpatient hospital stay is essential to have a proper IV identification strategy. Note that previous works using ICU congestion as an instrument (e.g., Kc and Terwiesch (2012), Shmueli et al. (2004)) were not able to account for the congestion during the patient's hospital stay.

Another mechanism that could invalidate the use of $ICUBusy_i$ as an IV is when periods of high congestion coincide with the arrival of very severe patients; this is what happens, for example, during an epidemic or a major accident affecting a large portion of the hospital's patient population. We tested this potential mechanism by analyzing the relationship between hospital occupancy and the LAPS score, a validated measure of patient severity, and found no correlation between the two. Although this does not prove that the instrument $ICUBusy_i$ is uncorrelated with the *unobservable* factors affecting outcomes, there is no reason to believe that they would be related to occupancy given that reasonable observable proxies of severity are not (this approach was also used by Kc and Terwiesch (2012) to validate a similar instrument).

Overall, our analysis provides substantial support validating the use of $ICUBusy_i$ as an IV. With this IV approach, the identification is driven by comparing differences in outcomes among patients who have similar observable characteristics captured by X_i but received different treatments because of the different levels of ICU occupancy at the time of their admission to an inpatient unit. Although this is not a perfectly randomized experiment, this identification strategy provides a valid approach to estimate the effect of ICU admission on patient outcomes.

In addition to $ICUBusy$, we consider other instrumental variables that were suggested as potential factors affecting ICU admission decisions from our conversations with nurses, physicians and hospital management. We refer to them as the set of *behavioral factors*. The first factor, $RecentDischarge_i$, accounts for recent discharges from the ICU and is motivated by the following mechanism. ICU discharges typically release the nurse who has been monitoring the discharged patient. The intensivist in charge may have an incentive to "preserve the nurse hours" by demonstrating a continuous demand for those nurses even after patients are discharged[6], leading to higher

[6]This behavior is related to supply-sensitive demand that has been shown in the medical literature. For instance,

ICU admission rates right after one or more ICU discharges. Note that this behavior is different from the speed-up effect reported in Kc and Terwiesch (2009) because it can also be manifested when discharges are not "forced" to occur faster. It is also different from the ICU occupancy effect because it can operate when the ICU has low utilization. To measure $RecentDischarge_i$, we count the number of all ICU discharges in the 3-hr window before patient i's admission to the first inpatient unit. In the sample, 56% of the patients see no recent ICU discharges, 27% see one discharge, and 11% see two discharges. Because bigger ICUs would naturally have more recent discharges, we divide the number of recent ICU discharges by the ICU capacity of each hospital to use it as $RecentDischarge_i$.

The second behavioral factor, $RecentAdmission_i$, accounts for the number of recent admissions of ED patients to the ICU. Since ICU beds are shared between ED and elective patients, a high number of recently admitted ED patients may reduce the bargaining power of the ED physician in his negotiation with the intensivist. To measure $RecentAdmission_i$, we consider ICU admissions in the 2-hr window before patient i's admission to the first inpatient unit, but count as a recent admission only if the patient is admitted via the ED to a medical service (excluding those that go to surgery, as in that case the negotiation may involve the surgeon). Because of shift changes, we do not expect the impact of expending negotiation power to propagate for extended periods of time. In our data, 84% of the patients see no recent admission and 14% see one recent admission. Similar to $RecentDischarge_i$, we divide the number of recent admissions by the ICU capacity of each hospital to define $RecentAdmission_i$. The third behavioral factor, $LastAdmitSeverity_i$, measures the severity of the last patient admitted to the ICU from the ED. The motivation for including this variable is that the most recent admit serves as a reference point in the negotiation process. If the ED physician just treated a very severe patient, he might require a new patient to be also very sick to recommend ICU admission. We define $LastAdmitSeverity_i$ as a dummy variable indicating whether the last patient admitted to the ICU had a LAPS score greater

see Wennberg et al. (2002) and Baker et al. (2008).

than or equal to the 66^{th} percentile value of the observed LAPS distribution. Table 4.2 provides summary statistics of the covariates for all the patients in our sample as well as patient grouped by whether they were admitted to the ICU or not.

Table 4.2: Summary statistics of patient characteristics, grouped by whether their first inpatient unit was an ICU versus non-ICU bed

	Non-ICU	ICU	ALL
Num. of obs.	63197	6936	70133
Selected X Covariates			
Age	67.3 (17.8)	64.0 (18.0)	67.0 (17.8)
LAPS	23.5 (18.1)	36.1 (25.2)	24.7 (19.3)
$\hat{P}(Mortality)$	0.044 (0.067)	0.095 (0.131)	0.049 (0.077)
Female	0.546	0.495	0.541
Z Covariates			
$ICUBusy$	0.096	0.039	0.091
$RecentDischarge$	0.033 (0.048)	0.040 (0.052)	0.034 (0.049)
$RecentAdmission$	0.009 (0.022)	0.009 (0.021)	0.009 (0.022)
$LastAdmitSeverity$	0.341	0.311	0.338

Note. Average and standard deviation (in parentheses for continuous variables) are reported.

The behavioral factors – $RecentDischarge_i$, $RecentAdmission_i$ and $LastAdmitSeverity_i$ – exhibit no correlation with the LAPS score of the incoming patient, suggesting that they are unrelated to patient severity and therefore appear to be exogenous. This is expected given the randomness in the arrival process of new incoming ED patients.

We define the vector of IVs, labeled Z, as these three behavioral factors plus $ICUBusy$. The next section describes how to implement the estimation using these IVs to instrument for the endogenous variable $Admit_i$.

4.4.3 Estimation

When the patient outcome is modeled via a linear regression as in (4.1), we can use a standard two stage least squares (2SLS) approach to implement the IV estimation. But because admission

decisions and some of our patient outcomes are discrete, a more efficient estimation approach is to develop non-linear parametric models to characterize the admission decision and the patient outcome and estimate these two models jointly via Full Maximum Likelihood Estimation (FMLE) (Wooldridge 2010). We describe this approach next.

The ICU admission decision is binary and is modeled through a Probit model defined by:

$$Admit_i = \begin{cases} \text{admit to ICU} & \text{if } X_i\theta - Z_i\alpha + \xi_i \geq 0, \\ \text{re-route to Ward} & \text{otherwise.} \end{cases} \tag{4.2}$$

where X_i are observable patient characteristics, Z_i are the IVs and ξ_i is an error term following a Standard Normal distribution.

Patient outcomes are modeled using two different approaches depending on whether the outcome is measured as a binary or a counting variable. We first consider the three binary patient outcomes $Mortality$, $TransferUp$ and $Readmit$. To model each of these outcomes, we use a Probit model defined by a latent variable:

$$\begin{aligned} y_i^* &= \beta_1 Admit_i + X_i\beta_2 + \omega_{h(i)} + \beta_3 AvgOccVisited_i + \varepsilon_i \\ y_i &= \mathbb{1}\left\{y_i^* > 0\right\}, \end{aligned} \tag{4.3}$$

where y_i^* is the latent variable. The additional control $AvgOccVisited_i$ captures the effect of the congestion during the hospital stay of the patient, as previously discussed. To account for the endogeneity in ICU admission decisions $Admit_i$, we allow for the error term ε_i to be correlated with the unobservable factors affecting admission (ξ_i in equation (4.2)) by assuming that the random vector (ξ_i, ε_i) follows a Standard Bivariate Normal distribution with correlation coefficient ρ (to be estimated along with the other parameters of the model). Note that this requires a joint estimation of the ICU admission model (4.2) and the outcome model (4.3). The model becomes a Bivariate

Probit which can be estimated via the Full Maximum Likelihood Estimation (FMLE) (Cameron and Trivedi 1998). The endogeneity of the admission decision $Admit_i$ can be tested through a likelihood ratio test of the correlation coefficient ρ being different from zero.

The patient outcome defined by $HospLOS_i$ is a count variable of the number of nights a patient stays in the hospital. A Poisson model could be used to model this count variable, but preliminary analysis of $HospLOS_i$ reveals over-dispersion (Table 4.3 shows the mean of $HospLOS_i$ is 3.9 while the variance is 24.0). Hence, we use the Negative Binomial regression, which can model over-dispersion using the parametrization developed in Cameron and Trivedi (1986). We use the extension developed by Deb and Trivedi (2006) to include a binary endogenous variable – the ICU admission decision $Admit_i$ – into the negative binomial regression, which is estimated jointly with model (4.2). The negative binomial regression includes the same covariates as in (4.3). The next section describes the estimation results of all the outcome models.

Table 4.3: Summary statistics of the patient outcomes

Outcome	n	Mean	Standard deviation	Median
Mortality	70,133	0.04	-	-
TransferUp	68,200	0.03	-	-
Readmission - 2 weeks	67,087	0.10	-	-
Hospital LOS (days)	70,133	3.9	4.9	3.0

4.5 Estimation Results

In this section, we discuss the results of the patient outcome models, which are summarized in Table 4.4. As discussed in Section 4.4.3, we estimate the admission decision and patient outcome model jointly to account for the endogeneity of the admission decisions. We find that all of our instruments have an impact on whether a patient is admitted to the ICU. For example, we find that when the ICU is busy, the likelihood of being admitted to the ICU decreases by 53% on average (statistically significant at the 0.1%). For space limitations, Table 4.4 shows only the coefficient and

the marginal effects of $Admit_i$ (i.e., whether the patient was admitted to the ICU or not), which is the main focus of this analysis. Each row corresponds to a different outcome (the dependent variable).

Table 4.4: Estimation results of the effect of ICU admission on patient outcomes

	With IV					Without IV
Outcome	Estimate (SE)	AME	ARC	ρ (SE)	Test $\rho = 0$	Estimate (SE)
Mortality	0.01	0.001	+1.6%	0.20**	0.00	0.42***
	(0.13)			(0.07)		(0.03)
Readmit	-0.22^{+}	-0.034	-32.2%	0.15*	0.03	0.05*
	(0.13)			(0.07)		(0.02)
TransferUp	-0.65***	-0.028	-77.3%	0.32**	0.00	-0.08*
	(0.16)			(0.10)		(0.04)
HospLOS	-0.44***	-1.2	-33.0%	0.56***	0.00	0.28***
(days)	(0.01)			(0.01)		(0.01)

Note. Each row corresponds to a different outcome (the dependent variable); AME - Average Marginal Effect; ARC - Average Relative Change; Standard errors in parentheses. $^{+}(p < 0.1), *(p < 0.05), **(p < 0.01), ***(p < 0.001)$.

In Table 4.4, the coefficients of $Admit_i$ are negative and significant in all models except $Mortality$, suggesting that admitting a patient to the ICU reduces the chance of having an adverse outcome. (Later we discuss possible explanations for the lack of significance in the $Mortality$ outcome model). The table also displays the average marginal effect (AME), which is the average expected absolute change in the outcome (among all patients) when a patient is admitted to the ICU instead of the Ward. The average relative change (ARC) is also reported, which is AME divided by the average outcome when a patient is not admitted to the ICU. The magnitude of the effect is substantial. For instance, admitting a patient to the ICU reduces the likelihood of hospital readmission by 32% on average.

The column "Test $\rho = 0$" shows the p-values of the test with the null hypothesis of exogeneity of the ICU admission decision, which is equivalent to a likelihood ratio test against the model where the correlation coefficient between the admission and outcome models' errors, ρ, is restricted to be zero. The estimates of ρ are reported in the column "ρ (SE)." In all models, the null hypothesis of exogeneity of the ICU admission decision is strongly rejected. Hence, the results suggest that

accounting for the endogeneity of the ICU admission decision is important to obtain consistent estimates of the effect of ICU care on patient outcomes.

We now assess the magnitude of the bias induced by neglecting the endogeneity of the admission decision in the estimation. The right panel of Table 4.4 ('Without IV") shows the estimates ignoring the endogeneity of the admission decision, which are significantly different from those estimated with IVs (left panel). All cases exhibit positive biases on the coefficients when ignoring the admission decision endogeneity. This is consistent with the endogeneity problem discussed in Figure 4.4. ICU patients tend to be more severe, and because part of the patient severity is unobserved and therefore cannot be controlled for, the naive estimates (without IVs) tend to underestimate the benefit of ICU admission. In some cases the bias is so severe that it leads to a positive correlation between being admitted to an ICU and experiencing adverse outcomes.

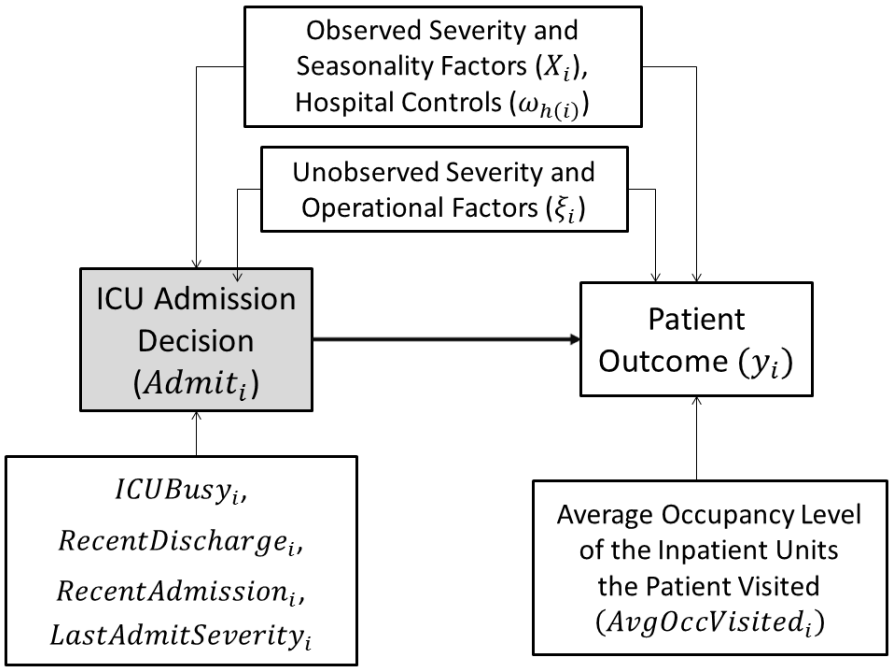


Figure 4.4: Relationship between ICU admission decision, patient outcome and observed/unobserved patient severity. The instrumental variables used to account for the endogeneity of the admission decision ($Admit_i$) are shown in the bottom-left box.

In all of our estimates, we could not find a significant effect of ICU admission on mortality

rates, which was at first surprising given the magnitude of the effect for other outcomes. A possible explanation of this relates to the IV estimation approach when the effects on the outcome are heterogeneous across patients. The estimation with valid IVs provides an unbiased effect of the average effect of ICU admission on patient outcomes *over the subset of patients that are affected by the instrument*. In our context, this includes patients whose ICU admission decision was affected by the ICU congestion one hour prior to their ED discharge. Figure 4.4.2 shows that this set includes patients from a broad class of severity – the ICU admission rate drops significantly when the ICU is congested and this is observed for patients from all severity classes. However, anecdotal evidence from our conversations with physicians in this hospital network suggest that if a patient is at high risk of death and ICU care and monitoring could substantially reduce these risks, ICU congestion is unlikely to have much effect on the patient's admission to the ICU. [7] This suggests that ICU congestion plays no significant role in determining the admission decisions of patients with true risk of dying. Therefore, our estimation approach cannot be used to measure the benefit of ICU admission for this subset of patients as they do not comply with the instrumental variable.

4.5.1 Robustness Analysis and Alternative Model Specifications

This section describes analyses using alternative specifications that support the robustness of our main results. Some of the controls of patient severity–LAPS and $\hat{P}$(Mortality)–are included with piece-wise linear functions to account for their possible non-linear effects on admission decisions and patient outcomes. We tried different specifications of these functions and the results were similar.

In the ICU admission model, we tested alternative measures to capture the level of occupancy in the ICU. As discussed in Section 4.4.2, our data analysis suggests that most of the adjustment to the ICU admission rate occurs when ICU occupancy goes above the 95^{th} percentile; hence,

[7] This gets more complicated by the patients who are denied ICU admission because they are deemed "too sick for ICU treatment" or have executed Do-Not-Resuscitate (DNR) orders; e.g., see Reignier et al. (2008).

$ICUBusy_i$ was defined as a binary variable indicating occupancy levels above this threshold. This measure accounts for the differences in ICU sizes across the hospitals in the sample. In addition, we tested other specifications in which we interact several hospital characteristics with $ICUBusy_i$ to account for potential heterogeneous effects: these included measures of hospital size (dividing hospitals into groups by size), the presence of an intermediate care unit at the hospital, as well as with different shifts (7am-3pm, 3pm-11pm, and 11pm-7am). In all cases the estimated average effect of ICU occupancy on ICU admissions was similar to what was obtained in the main results.

In our model, we control for month of admission to capture potential seasonal effects and also hospital fixed effects to account for variations across hospitals. It is possible that there are time-varying hospital characteristics, which would not be controlled for with our month and hospital fixed effects. Thus, we also included hospital-month fixed effects and found that while these effects do seem to be statistically significant, accounting for them does not change our main results.

In defining $RecentDischarge_i$ and $RecentAdmission_i$ in the ICU admission model, we use the 3-hr and 2-hr time windows, respectively. We experimented with shorter and longer time windows. For $RecentDischarge_i$, we observed that the effect persisted even when we consider a 8-hr time window (which we consider as the maximum duration since shifts change every eight hours). For $RecentAdmission_i$, increasing the time window gave us weaker results, and the effect of this variable disappeared when we considered time windows longer than three hours. The estimates of the other model coefficients were robust to these alternative specifications.

Furthermore, we observe that the behavioral factors are less powerful IVs than $ICUBusy_i$, in the sense that they explain less variation in the ICU admission decision. We also considered specifications that had $ICUBusy_i$ alone as an IV and the results were similar.

We also examined other factors which may affect the admission decision, such as the severity of the patients currently in the ICU. Because our measures of severity are taken at the time of hospital admission (not at the time of ICU admission or any time later in their hospital visit), this measure may not be very accurate, especially as we cannot account for how patient severity

improves or deteriorates during their ICU stay. Nonetheless, when we control for the average severity of patients in the ICU, we find that 1) a patient is less likely to be admitted to the ICU when there are many severe patients and 2) the main results (e.g. impact of a busy ICU on admission and the effect of admission on outcomes) of our estimations are robust to these alternate specifications.

We use the Full Maximum Likelihood Estimation (FMLE) to estimate the patient outcome models. While being more efficient, the FMLE imposes strong parametric assumptions on the distribution of outcomes. We did some validation of these assumptions for the count variable $HospLOS$. We observe over-dispersion–the unconditional variance is 24.0 while the mean value is 3.9–and no evidence of zero-inflation–only 5.9% had hospital LOS equal to 0. Hence, the negative binomial model seems an appropriate model for this outcome.

For $Readmit$, recall that we have set the time window of two weeks after discussions with doctors. We have tested shorter and longer time windows and the results for the two week time window demonstrated higher statistical significance and magnitude.

All the outcome models include the covariate $AvgOccVisited_i$ to control for the average occupancy level during each patient's stay in the hospital. We considered other alternatives to measure the effect of this factor: (i) the daily average occupancy of all the inpatient units in the hospital during the patient's hospital stay; (ii) the maximum occupancy level experienced by the patient in an inpatient unit during his hospital stay; (iii) the average number of inpatients in the hospital during the patient's hospital stay over the maximum possible number of inpatients (without differentiating amongst different inpatient units); and (iv) the average occupancy level of inpatient units at the time the patient was discharged from the first inpatient unit he visited. All of these alternative definitions gave results that were consistent with what we report for our main specification.

When analyzing $TransferUp$, we included all patients in the estimation model as long as the patient had been to a non-ICU at least once. But patients who had in-hospital death may have a lower probability of a transfer-up event. Hence, we excluded patients with in-hospital death in $TransferUp$ model and found that the results were similar.

For the $HospLOS$ model, recall that we measured it by the number of nights a patient stayed in the hospital after being discharged from the ED. We tried defining $HospLOS$ as LOS rounded to the nearest day, and the results were similar. We also estimated the outcome models excluding patients with in-hospital death for the $HospLOS$ model, and the results were again similar.

4.5.2 Accounting for Alternative Mechanisms that Control ICU Congestion

Although the results seem to be robust to alternative specifications, it is possible that the effect we attribute to ICU admission may be in part capturing the effect of other mechanisms used by the hospitals to manage ICU capacity. In this section, we consider two such alternative mechanisms.

The first mechanism, which has been studied by Anderson et al. (2011) and Kc and Terwiesch (2012), is to shorten or "speed-up" the time a patient stays in the ICU to make room for new severe patients. Kc and Terwiesch (2012) show that this speed-up increases the probability of readmission of those patients, which is one of the patient outcomes we analyze in this study. Because this mechanism is more likely to be used when the ICU is busy, it is correlated with our main IV and can therefore confound our estimation of the effect of ICU admission on patient outcomes.

The speed-up effect analyzed in Kc and Terwiesch (2009) was based on cardiac surgery patients, whereas our study is based on ED patients, a completely different patient population. Therefore, we replicate their methodology in our patient sample to measure the magnitude of the effect. To further validate the replication of this methodology, we estimated the same model using a sample of patients comparable to the one studied in Kc and Terwiesch (2012): we utilized our data that also include elective surgical patients admitted to the ICU. We define *firstICU LOS*$_i$ as the ICU length of stay during patient i's first ICU visit and $BUSY_i$ as the bed utilization of the ICU at the time patient i was discharged from this ICU visit. Because our dataset does not have information on the number of scheduled arrivals, our definition of $BUSY_i$ is not the same as in Kc and Terwiesch (2012). Instead, we let $BUSY_i$ be 1 if the number of existing ICU patients at the time patient

i is discharged from the ICU exceeds the 95^{th} percentile of occupancy.[8] We estimate the effect of ICU occupancy on ICU length of stay through the following regression:

$$log(firstICU\ LOS_i) = \gamma BUSY_i + \beta X_i + u_i, \tag{4.4}$$

where X_i is a vector of observable patient characteristics that describe the patient's severity of illness. A negative γ suggests that high ICU congestion leads to a shorter ICU LOS – a speed-up effect.

Table 4.5: Estimation Results of the speed-up model in (4.4)

	$Busy$ Coefficient (Standard Error)	# Observations	R^2
ED, Medical	-0.02 (0.03)	10521	0.16
Non-ED, Surgical	-0.13** (0.04)	4524	0.14

Note. $^{+}(p < 0.1), *(p < 0.05), **(p < 0.01), ***(p < 0.001)$.

The regression model (4.4) is estimated with two samples of patients that were admitted to the ICU: (1) ED, medical patients which is this study's main cohort and (2) elective surgery patients. The estimation results are reported in Table 4.5. The results of this analysis cannot reject the null hypothesis of no speed-up effect in our patient population (p-value 0.47), but they strongly reject the null hypothesis of no speed-up effect (p-value 0.001): a congested ICU reduces length of stay in the ICU by 12%. Therefore, our method correctly replicates the results of Kc and Terwiesch (2012), but at the same time shows no speed-up effect in the patients admitted to the ICU via the ED. We conclude that this mechanism is not relevant in our patient population and therefore cannot be confounding our main results regarding the effect of ICU admission on patient outcomes.

It is also interesting to see how the mechanisms to manage ICU capacity may vary across patient types. This was also reported in Chen et al. (2013), showing that in contrast to non-cardiac patients, severity scores have little impact on the admission decision for cardiac patients.

[8] We have tried various specifications for defining $BUSY$, such as using different cutoff points for occupancy level and including future arrivals in a certain time window, and the results were consistent. In addition, we have tried hazard rate models–Weibull and Cox proportional hazard models–with $BUSY$ measure included as both time-invariant and time-varying, and the results were consistent.

The second mechanism is ED boarding: a congested ICU can extend the time a patient spends in the ED waiting to be transported to an inpatient unit. ED boarding – patients waiting in the ED to be admitted to an inpatient unit – tends to increase when the inpatient unit where the patient was admitted to is more congested. Hence, patients who are admitted to the ICU during high periods of ICU congestion may have waited a longer time in the ED. Since the ED has less adequate resources to take care of the patient, this additional waiting time in the ED may have direct implications on the patient outcome.[9] This suggests that ICU congestion may influence patient outcomes through two different mechanisms: (i) the ICU admission decision, which is captured through model (4.2) and; (ii) the ED boarding time. Consequently, for ICU congestion to be a valid instrumental variable in isolating the effect of ICU admission on patient outcomes, we need to control for the effect of ED boarding time in the outcome model.

To account for this mechanism, we include a measure of ED boarding time as a covariate in the outcome models (equations (4.1) and (4.3)). ED boarding time is defined as the time between the decision to admit the patient until the patient is discharged from the ED and physically moved to the inpatient unit, which is measured in the data. If a patient's ICU admission has been delayed (shown by long ED boarding time), the patient's outcomes might be adversely affected by not receiving timely care. Therefore, the effect of ED boarding time should be negative. However, ED boarding time is endogenous and can be affected by unobservable patient characteristics related to the patient's outcome. A severe patient that requires urgent care is likely to have a shorter boarding time.

For this analysis, ED boarding time ($EDboard$), defined as the time between the decision to hospitalize the patient until the patient is discharged from the ED and physically moved to the inpatient unit, is added as an additional covariate in the outcome models (4.1) and (4.3). This new specification has two endogenous covariates: $Admit$ and $EDboard$; the former is instrumented

[9]California requires 1:3 nurse-to-patient ratio for EDs, which is lower than that of ICUs but higher than that of general wards. Moreover, the primary purpose of an ED is to stabilize patients, rather than to provide supportive care as given in inpatient units.

by $ICUBusy$ so we need additional instruments for the latter. A valid exogenous instrumental variable affects ED boarding time but is unrelated to the severity of the patient. The instrument we use is the "average level of bed occupancy of the inpatient unit the patient goes to after the ED", labeled $FirstInpatientOcc$, where the average is taken during the time the patient is boarding in the ED. The logic is similar to our $ICUBusy$ instrument: if the patient was routed to an inpatient unit but this unit was busy when the patient was in the ED, then the patient probably had to stay a longer time in the ED waiting for a bed. Recall that $ICUBusy$ is based on the level of occupancy of the ICU one hour prior ED discharge, whereas $FirstInpatientOcc$ measures the occupancy of ICU or the ward, depending on where the patient is routed to after the ED. Hence, the two instrumental variables are not perfectly correlated. A regression of the logarithm of ED boarding time on $FirstInpatientOcc$ shows a positive and highly significant effect; a 10% increase in the inpatient occupancy increases ED boarding time by 18%. For this model, we use similar controls as in our earlier specification. Details of the regression output is available from the authors upon request.

The estimation of the model goes as follows. Since the outcome models are not linear, we use a control function approach to implement this IV estimation. The estimation is carried out in two steps: (1) we first estimate a linear regression with *log(ED Board)* as the dependent variable and the IVs and controls as covariates; and (ii) we calculate the residuals of this regression and include the residuals and *log(ED Board)* as additional covariates in the outcome model. See Wooldridge (2010) for more details on the control function approach.

This econometric model identifies the effect of ED boarding and ICU admissions on patient outcomes, partialling out the effect of each variable separately; that is, it measures the effect of ICU admission above and beyond any effect caused by ED boarding. Table 4.6 reports the estimated coefficients for ICU admission and *log(ED Boarding Time)* for the different outcome models. They show that, for some outcomes, a longer ED boarding time leads to worse patient outcomes; for other outcomes the effect is not significant. More importantly, the estimated effects of ICU admis-

Table 4.6: Estimation results of the patient outcome model including ED boarding time as an endogenous covariate

Outcome	ICU admission	Log(ED board)
Mortality	0.03 (0.13)	0.05 (0.04)
Readmit	-0.21 (0.13)	-0.01 (0.03)
TransferUp	-0.61*** (0.16)	0.16*** (0.04)
HospLOS (days)	-0.40*** (0.01)	0.01 (0.01)

Note. Standard errors in parentheses. $^{+}(p < 0.1), *(p < 0.05), **(p < 0.01), ***(p < 0.001)$.

sion are similar to those reported in Section 4.5. The main conclusion of this analysis is that our main results regarding the effect ICU admission on patient outcomes are not confounded by the effect of ED boarding time.

4.6 Evaluating Alternative Admission Policies

A primary objective in our study is to quantify the benefits of ICU care. This is an essential first step in comparing different ICU admission strategies. To examine how we can utilize the measures we have just estimated, we consider a parsimonious model of patient flows into the ICU to examine the performance of various admission policies. We leverage our estimation results to calibrate a simulation model, which allows us to compare patient outcomes across different admission policies. In particular, we are interested in studying whether admission criteria that are based on objective metrics of patient risk can outperform the current hospital admission policies.

4.6.1 Model of Admission Control

e model the ICU admission control problem as a discrete version of the Erlang loss model, similar to the one used in Shmueli et al. (2003). This admission control problem can be viewed as a special case of the stochastic knapsack problem studied in Altman et al. (2001), and we leverage some results from that work to characterize its solution.

Consider an ICU with B beds. In order to focus on the ICU admission decision, we assume there is ample space in the other inpatient units to care for all patients. We denote by x the number of occupied ICU beds at any given point in time. When $x = B$, arriving patients must be routed to the general Ward. Time is discretized into periods of fixed length dt, indexed by t, where the periods are sufficiently short so that it is reasonable to assume at most one patient arrives in a given period. In each period, a patient arrives to the ICU with probability λ. Upon arrival, a decision must be made on whether to admit the patient or not. If admitted to the ICU, a patient's length of stay is geometrically distributed with mean $1/\mu$. We assume that patient discharge is exogenous, i.e. there is no speed-up in the ICU.[10]

If a patient is routed to the Ward an expected cost of ϕ_c is incurred, where c indexes the customer's class. Without loss of generality, classes are numbered $1 \dots C$ so that ϕ_c increases in c; therefore, classes can be interpreted as the severity of the patient, where the benefit of admitting a patient increases with his severity. The objective is to choose an admission criteria that minimizes the total expected cost over a finite-horizon.

An *admission policy* is defined as a decision rule that chooses whether to admit or reroute an incoming patient, for each possible state characterized by the class of the incoming patient (c) and the number of occupied ICU beds $x \in [0, B]$. Altman et al. (2001) shows that the optimal admission policy is a threshold policy with the following structure. Given a occupancy level x, admit a patient if and only if his class satisfies $\phi_c \geq \kappa_x$. The values $\{\kappa_1 \dots \kappa_B\}$ are referred to as the *optimal thresholds*. It is also shown that the thresholds κ_x are increasing in x.

Next, we describe how we set the primitives of this admission control problem in order to run a simulation.

[10] As discussed in Section 4.2, other mechanisms may be used; although we do not find that speedup is used for the patient group we study; see Section 4.5.2. Via numerical analysis, we found that the qualitative results extend when speed-ups are incorporated.

4.6.2 Model Calibration and Simulation

The simulation analysis focuses in an ICU with $B = 21$ beds, which is the median ICU size in our data. To simulate ICU admissions, we sample (with replacement) patient characteristics from a hospital whose 95^{th} percentile of occupancy distribution was at 20 beds and 99^{th} percentile at 21 beds. This hospital treated 7,387 ED-medical patients during our study period. Each discrete time period lasts 10 minutes and patients arrives to the ICU with probability λ so that on average 3 patients arrive per hour. This has been delicately chosen so that the simulated setting is consistent with the regime of the hospitals in our study that admit approximately 10% of the inpatients to the ICU under the current policy. The average patient LOS in the ICU is $1/\mu = 60$ hours, which corresponds to the average duration of ICU stay in our sample.

Next, we describe how to estimate the expected rerouting costs ϕ_c for each patient class c. This requires defining the health outcome measure to be considered–$HospLOS$, $TransferUp$, and $Readmit$ (we do not study mortality since the estimates for that outcome were imprecise and not statistically significant). Let y be the outcome of interest. Recall that ϕ_c represents the difference in this expected health outcome if a patient is admitted to the ICU versus not admitted.

Information about the incoming patient is essential to assess his severity class. Each patient i is fully described by a set of objective characteristics X_i (recorded in our data and described in Table 4.2) and the "error term" ξ_i capturing other patient characteristics, not observed in the data, that are taken into account by the physician when assessing the patient admission. We therefore call X_i and ξ_i the objective and discretionary component of the patient information, respectively. Defining an admission policy requires specifying what kind of information is considered when making a decision, which we define as the information set I_i. We focus on studying policies that use all the information, $I_i = (X_i, \xi_i)$, and policies that use only the objective component, $I_i = X_i$.

For a given patient with information set I_i, the expected rerouting cost is calculated as:

$$\phi_i = E(y_i|Admit_i = 0, I_i) - E(y_i|Admit_i = 1, I_i), \tag{4.5}$$

where the expectation is taken with respect to ε_i, the error term in the corresponding outcome model. We explain in detail how we estimate this cost for $Readmit$ with information set $I_i = X_i$; the calculation for the other metrics is similar. For readmissions, Equation (4.5) becomes:

$$\phi_i^{Readmit} = \Pr(\varepsilon_i \geq -\beta_2 X_i) - \Pr(\varepsilon_i \geq -\beta_2 X_i - \beta_1),$$

which is positive when $\beta_1 < 0$. When only the objective information component is observed, ε_i follows a Standard Normal distribution. When the discretionary component ξ_i is also included in the information set (i.e., $I_i = (X_i, \xi_i)$), ε_i follows a Normal distribution with mean $\rho\xi_i$ and variance $(1-\rho)^2$. The parameters β_1, β_2, ρ are the estimates of the Readmission outcome reported in Section 4.5 and therefore the probabilities can be calculated numerically.

Equation (4.5) calculates the rerouting cost for a specific patient. In practice, deriving the optimal admission policy via dynamic programming requires a finite set of patient classes. To achieve this for each health outcome, we first calculate ϕ_i for all 7,387 patients treated in the hospital that we chose to simulate. (For the case where the discretionary component ξ_i is included in the information set and, hence, the value of (4.5) depends on ξ_i, we generate 1,000 realizations of ξ_i and compute 7,387,000 values of ϕ_i. Then we partition patients into 10 groups based on the deciles of this distribution; each patient class has lower and upper bound on ϕ_i which defines patients that belong to the class group. Class c's rerouting cost ϕ_c is set as the average rerouting costs of the patients in that class.

A *policy* is specified by a function that maps patient information set I_i and the number of occupied beds (x) to an admission decision. The following procedure describes how we carry out our discrete-time simulation of a given policy. At $t = 0$, occupancy is set to zero. In every period, with probability λ, a patient is sampled from the population of patients, characterized by X_i and a random vector (ξ_i, ε_i) from a Bivariate Standard Normal with correlation coefficient ρ. A patient is admitted to the ICU if $x < B$ and the policy evaluates to do so. This will result in an increase

in ICU occupancy to $x + 1$. Otherwise, the patient is not admitted. At the end of the period, each ICU patient leaves with probability μ. We simulate a full year, with one month of warmup after which the system status reaches stationarity, over 1,000 iterations.

4.6.3 Admission Control Policies

We use the simulation model described above to examine how different ICU admission strategies impact aggregate patient outcomes. We compare 4 different policies. The **Estimated Current Policy** corresponds to an empirical model of the admission policy used at the hospitals in our study, which we estimate from the data. The **Optimal Objective Policy** uses the objective component of patient information (i.e. $I_i = X_i$) to assess the expected rerouting cost, and derive the optimal threshold levels of admission. The **Optimal Full Policy** uses the objective and discretionary components ($I_i = (X_i, \xi_i)$) in assessing the expected rerouting cost. The fourth policy is similar to the Estimated Current Policy, but with $B = 22$ as bed capacity. We now describe each of these policies in more detail.

Estimated Current Policy: The structural results of Altman et al. (2001) establish that the optimal policy is of threshold form. Although the policy currently used by the hospital need not be optimal, Figure 4.4.2 presents several patterns that are consistent with a threshold policy. First, admission rates tend to increase as patient severity increases. Second, admission rates decrease at higher levels of occupancy, consistent with threshold levels that increase with the number of occupied beds. Third, the drop in admission rate due to an increase in occupancy is higher for more severe patients, which can be shown to be in line with a threshold policy.[11]

We restrict the hospital we choose to simulate to follow a threshold policy that uses an infor-

[11]Consider two patient classes, high (H) and low (L) severity, and assume that patient severity for class $j \in \{L, H\}$ follows a Normal(μ_j, σ^2), where $\mu_L < \mu_H$. Given a threshold κ, the admission probability for patient class j is given by $\Pr(N(\mu_j, \sigma^2) > \kappa)$; assume $\mu_L < \mu_L < \kappa$ (less than 50% of patients in all classes are admitted). An increase in occupancy raises the threshold to $\kappa + \Delta$, which decreases the admission rates of all groups, but the H group decreases by more. These results are not specific to the Normal distribution assumption on severity – they hold for any distribution with density function decreasing at the threshold κ (i.e. $f'(x) < 0$ for $x > \kappa$).

mation set $I_i = (X_i, \xi_i)$ and develop an empirical model to estimate the parameters of this policy. The model is given by:

$$Admit_i(I_i, x) = 1\{X_i\theta + \xi_i \geq f(x;\kappa)\}, \tag{4.6}$$

where $f(x;\kappa)$ is a function that parameterizes the thresholds as a function of the occupancy x. Assuming $\xi_i \sim N(0,1)$, the model can be estimated via a Probit model. We experiment (and hence fit the Probit model) with all possible combinations of the way the occupancy x can affect the admission policy; that is, we vary the number of thresholds and the locations of the thresholds that the occupancy x can have. For instances, $f(x;\kappa)$ can change at every possible occupancy level or it can change only once, say when the ICU occupancy is 20 and above. For each model (4.6) with different combination for $f(x;\kappa)$, we compute the Bayesian Information Criterion (BIC), which is a commonly used metric to select the most parsimonious model that best describes data; it is computed based on the likelihood and has a penalty term for the number of parameters in the model; see Raftery (1995). We then choose the model that has the smallest BIC value to be our estimated current policy.

Optimal Policies: Since the optimal policy is of threshold form, a patient i is admitted if:

$$Admit_i(I_i, x) = 1\{\phi_i > \kappa_x\},$$

where ϕ_i is calculated by equation (4.5). We use dynamic programming to determine the threshold values $\{\kappa_x\}_{x=0}^{B}$ that minimize total costs. Notice that the calculation of ϕ_i depends on the information set I_i, therefore the optimal policy depends on I_i, which leads to the **Optimal Objective Policy** ($I_i = X_i$) and the **Optimal Full Policy** ($I_i = (X_i, \xi_i)$). To facilitate the dynamic programming recursion, we assign patient i the rerouting cost of his class ϕ_c, which reduces the possible values of each threshold to $\{\phi_1 \dots \phi_{10}\}$. This provides a lower bound on the performance of the optimal policies.

The Estimated Current Policy may perform worse than the Optimal Objective Policy. This

could occur for several reasons. First, the admission decision under the optimal objective policy is based on the rerouting cost ϕ_i, while, in the current admission decision described by equation (4.6), the left hand side of the inequality is not necessarily equal to ϕ_i. More specifically, this implies that the estimated current policy may not be appropriately weighting the objective metrics X_i. Second, the threshold adjustment function $f(x;\kappa)$ may not set the optimal threshold levels that properly accounts for the opportunity cost of using up a bed, which the optimized policy can. On the other hand, the Estimated Current Policy has a richer information set than the Optimal Objective Policy, so it is not a priori known which will perform better. Because the Optimal Full Policy utilizes the same information as the Estimated Current Policy, accurately weights both the objective and discretionary information ($I_i = (X_i, \xi_i)$), and optimizes the thresholds, the Estimated Current Policy will have worse performance.

4.6.4 Results and Discussion

Table 4.7 summarizes the simulated patient outcomes–$HospLOS$, $TransferUp$, and $Readmit$–under the alternative policies we consider. Noting that the ICU admission decision is an inherently multi-objective problem, we also consider a combined outcome which considers the impact of ICU admission on total hospital days in the current inpatient stay as well as any potential subsequent hospital stay due to readmission. In particular, we convert each readmission into the average stay of 3.9 hospital days (see Table 4.3) and add this to $HospLOS$; we note that this is a conservative measure as readmitted patients are likely to stay longer in the hospital[12]. Finally, for comparison purposes, we convert hospital days into dollar amounts utilizing an estimate of $2,419 per hospital day as given in The Kaiser Family Foundation, statehealthfacts.org (2012).

The column labeled "BASE-21 Beds" under "Estimated Current Policy" lists the performance of the estimated current policy in a 21-bed ICU. Under the current policy (estimated as described

[12]We do not include (convert) $TransferUp$ into total hospital days because, while patients who are transferred up tend to have longer LOS, this is captured in the effect of $HospLOS$. To avoid double-counting, we only combine $Readmit$ and $HospLOS$

Table 4.7: Simulation results of alternative ICU admission control policies

	Estimated Current Policy		Optimal Objective	Optimal Full
	BASE - 21 Beds	22 Beds	for Each Outcome	for Each Outcome
# Readmissions	**2550**	-3.7	-26.6	-35.3
# Transfer-ups	**762.9**	-5.9	14.8	-38.6
HospLOS (years)	**245.6**	-0.4	-2.0	-9.0
Total HospLOS (years)	**272.9**	-0.4	-2.2	-9.2
(Estimated savings)		(-$ 0.4 m)	(-$ 1.9 m)	(-$ 8.1 m)

Note. The performance measures of the current policy are denoted in bold; all other results are changes from the performance of the estimated current policy.

above), there were on average 2550 hospital readmissions, 762.9 transfer-ups to the ICU from the general wards, and patients spent a total of 245.6 hospital bed years over the course of a year. We note that our simulation results were well aligned with what we observe in the data (reported in Tables 4.2 and 4.3): in our simulations, approximately 10% of the patients were admitted to the ICU, 11% of the patients experienced readmissions, and 3% experienced transfer-up events.

In the second column, labeled "22 Beds", we also report the change in performance of the estimated current policy when we increase the ICU capacity by one bed. Increasing the ICU bed capacity by one bed could be quite expensive; we roughly estimate this cost to be $0.8 million per year based on a $3,164 expense per ICU day (Aloe et al. 2009). In examining alternative admission policies, we will examine if some of the improvements in patient outcomes can be achieved without this high investment cost of increasing capacity.

The third column, labeled "Optimal Objective", provides the performance of the optimal policy based on objective measures alone, so that the information set $I_i = X_i$. Each row corresponds to a different policy optimized to minimize the corresponding outcome. Because the optimal objective policy optimizes the admission thresholds while also utilizing the direct relationship between the available information (X_i) and patient outcomes, it can sometimes do better than the estimated current policy. This is the case when we use the optimal occupancy-dependent threshold derived from the cost function for readmissions ($\phi_c^{Readmit}$) and hospital LOS ($\phi_c^{HospLOS}$); we observe 26.6 fewer readmissions and 2 fewer years of hospital LOS on average compared to the current policy. On the other hand, the estimated current policy may outperform the optimal objective policy since

it utilizes more information ($I_i = (X_i, \xi_i)$) and as we saw in Table 4.4, ξ_i captures a significant portion of unobserved factors that affect patient outcomes (as indicated by the correlation coefficient ρ). Indeed, the optimal objective policy aimed at minimizing $Transferup$ has *more* transfer-ups (15 more on average) compared to the estimated current policy. That said, this was not a systematic effect. We examined other hospitals and found that the optimal objective policy can outperform the current policy across all patient outcomes. These results suggest that the discretionary information can be useful, but optimizing the admission decision based solely on objective criteria can often result in better patient outcomes.

Finally, we further explore the benefits of incorporating the doctors' discretionary information in the admission decision. The last column, labeled "Optimal Full", uses the objective and discretionary information ($I_i = (X_i, \xi_i)$) available to the physicians, and further optimizes the admission thresholds. We see that by optimizing the thresholds, patient outcomes can be universally improved compared to the estimated current policy. The difference between "Optimal Full" and "Optimal Objective" captures the value of the physicians' discretion with respect to how much patient outcomes can be improved when also incorporating ξ_i in the information set. We can see the benefit of the physicians' assessment can be quite substantial, resulting in 8.4 fewer readmissions, 52.7 fewer transfer-ups, and 6.8 fewer patient years spent in the hospital. Moreover, these gains are orders of magnitude greater than what we achieve by adding an additional ICU bed, without incurring the costs of finding space and paying for such a structural change.

4.7 Conclusion

We have examined the impact of ICU congestion on a patient's care pathway and the subsequent effect on patient outcomes. We focused on medical patients who are admitted via the emergency department: a large patient cohort that comprises more than half of the patients admitted to the hospital. This is the first study to provide objective metrics that can be used by ED doctors and

intensivists to decide which patients to admit to the ICU from the ED. We empirically found that the ICU congestion can have a significant impact on ICU admission decisions and patient outcomes and provided systematic and quantitative measures of the benefit of ICU care on various patient outcomes. Furthermore, we provide a detailed characterization of the optimal ICU admission policy based on objective measures of patient severity and show how to compute these policies for different patient outcome measures using empirical data, dynamic programming and simulation methods. Via simulation experiments, we were able to compare the performance of admission policies based purely on objective criteria (calculated from our empirical estimation) vis-à-vis the performance of the current admission policies used by each hospital in our study. We showed that for certain outcome measures, using optimal policies based on objective metrics alone can outperform current hospital policies. For other outcome measures, we found that the discretionary criteria used by doctors is useful and can help improve system performance relative to the decision based solely on objective criteria. We believe this is the first work to study the impact of doctors' discretionary criteria on system performance in a healthcare setting.

From an estimation perspective, our instrumental variable approach can be extended to estimate the effect of other operational decisions. It is often the case that the effect of operational decisions on service outcomes is hard to estimate because of endogeneity bias. Our identification strategy of using operational and behavioral factors as instrumental variables and carefully controlling for factors that would invalidate the instrument can be further utilized in related questions. We believe the present work can be easily applied to study capacity allocation and the impact of the occupancy level of available resources in many other healthcare settings. For instance, the differentiated levels of care can be among different ICU units. Rather than having only one type of ICU, many hospitals have specialized ICUs such as cardiac, surgical and medical ICUs, and the level of nurse-to-patient ratios and level of treatment might differ. However, they are sometimes shared when the occupancy levels are high in some of these units. Our model can be applied to estimate how the admission control to these different types of ICUs are done and whether it has an impact on patient outcomes.

We acknowledge that our study has several limitations, which in turn suggests future research directions. First, our dataset is limited in that all hospitals belong to one healthcare organization and that the majority of the patients are insured via this same organization. It would be interesting to look at other types of hospitals, which would enable us to explore features such as the difference between paying and non-paying patients. Second, in Section 4.6.1, we introduce a stylized model of ICU admission with constant arrival rate of inpatients and constant departure rate of ICU patients. We believe it serves its role of giving us insights on the impact of operational and medical factors on ICU admission control. Possible extensions of this simulation model could incorporate time-varying arrival rates, departure rates that depends on patient severity, and readmissions to the ICU and to the hospital. We note that incorporating these features adds new analytic challenges and that it is an active area of ongoing research (e.g., see Feldman et al. (2008) and Yom-Tov and Mandelbaum (2014)). Third, there is a limitation of our empirical strategy due to the fact that an estimation based on instrumental variables provides an estimate of the average effect of the endogenous variable on the population that is affected by the instrument. In our context, our approach measures the average effect of ICU treatment on health outcomes for those patients whose ICU admission decision depends on the congestion of the ICU. This excludes two sets of patients: (1) Patients that are never admitted to the ICU, even if there is ample space in the ICU. This set of patients is probably the ones that benefit the least from ICU treatment. (2) Patients that are severe enough to be admitted to the ICU no matter how busy it is. These are usually the most severe patients and includes those patients with higher risk of dying. Hence, a limitation of our IV strategy is that we cannot estimate the benefit of ICU admission for all patients, but only for those for which the hospital's admission decision was affected by the level of congestion of the ICU, which probably excludes the most severe and the more healthy patients. Estimating the effect for these extreme cases would probably require a randomized experiment, which would be ethically questionable especially for the high severity group. Lastly, we hope to tease out and quantify the impact of the different adaptive mechanisms discussed in Sections 4.2 and 4.5.2–delays and boarding, speed-up,

admission control, surgery cancellation and blocking via ambulance diversion– in terms of patient outcomes and hospital costs, depending on patient admission types and diagnosis. Building an analytic model that includes the complex interplay between different adaptive mechanisms on patient outcomes might prove useful in developing decision support tools for ICU admission, discharge, and capacity planning.

Bibliography

Aksin, O., M. Armony, V. Mehrotra. 2007. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management* **16** 665–688. 1, 3.1

Alexopoulos, C., N. Argon, D. Goldsman, N. Steiger, G. Tokol, J. Wilson. 2007. Efficient computation of overlapping variance estimators for simulation. *INFORMS Journal on Computing* **19**(3) 314–327. 2.4

Alexopoulos, C., D. Goldsman. 2004. To batch or not to batch? *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **14**(1) 76–114. 2.1.2, 2.6.1

Allon, G., S. Deo, W. Lin. 2013. The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* **61**(3) 544–562. 4.2

Aloe, K., M. Ryan, L. Raffaniello, L. Williams. 2009. Creation of an intermediate respiratory care unit to decrease intensive care utilization. *Journal of Nursing Administration* **39**(11) 494–498. 4.6.4

Altman, E., T. Jiménez, G. Koole. 2001. On optimal call admission control in resource-sharing system. *Communications, IEEE Transactions on* **49**(9) 1659–1668. 4.2, 4.4.2, 4.6.1, 4.6.3

Anand, K., H. Mendelson. 1997. Information and organization for horizontal multimarket coordination. *Management Science* **43**(12) 1609–1627. 4.2

Anderson, D., C. Price, B. Golden, W. Jank, E. Wasil. 2011. Examining the discharge practices of surgeons at a large medical center. *Health Care Management Science* **14**(4) 1–10. 4.2, 4.5.2

Armony, M., S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-Tov. 2011. Patient flow in hospitals: a data-based queueing-science perspective. *Working paper, New York University*. 1, 3.1, 4.3.2, 5

Asaduzzaman, M., T. Chaussalet, N. Tobertson. 2010. A loss network model with overflow for capacity planning of a neonatal unit. *Annals of Operations Research* **178** 67–76. 3.1

Asmussen, S. 2003. *Applied probability and queues*, vol. 51. Springer. 2.4.2

Asmussen, S., P. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis: Algorithms and Analysis*, vol. 57. Springer. 2.1.2, 2.4.1

Avramidis, A., A. Deslauriers, P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* **50** 896–908. 3.1.2

Azoulay, E., F. Pochard, S. Chevret, C. Vinsonneau, M. Garrouste, Y. Cohen, M. Thuong, C. Paugam, C. Apperre, B. De Cagny, et al. 2001. Compliance with triage to intensive care recommendations. *Critical Care Medicine* **29**(11) 2132–2136. 4.2

Baccelli, F., P. Brémaud. 2003. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, vol. 26. Springer. 2.1

Baker, L., S. Atlas, C. Afendulis. 2008. Expanded use of imaging technology and the challenge of measuring value. *Health Affairs* **27**(6) 1467–1478. 6

Barbour, A., L. Holst, S. Janson. 1992. *Poisson Approximation*. Oxford University Press, Oxford, U. K. 3.1

Bassamboo, A., A. Zeevi. 2009. On a data-driven method for staffing large call centers. *Operations Research* **57**(3) 714–726. 3.1.2

Batt, R., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper, The Wharton School*. 4.2

Bertsimas, D., G. Mourtzinou. 1997. Transient laws of nonstationary queueing systems and their applications. *Queueing Systems* **25** 315–359. 1.1

Boumendil, A., D. Angus, A. Guitonneau, A. Menn, C. Ginsburg, K. Takun, A. Davido, R. Masmoudi, B. Doumenc, D. Pateron, et al. 2012. Variability of intensive care admission decisions for the very elderly. *PloS ONE* **7**(4) e34387. 4.1

Brilli, R., A. Spevetz, R. Branson, G. Campbell, H. Cohen, J. Dasta, M. Harvey, M. Kelley, K. Kelly,

M. Rudis, et al. 2001. Critical care delivery in the intensive care unit: defining clinical roles and the best practice model. *Critical Care Medicine* **29**(10) 2007–2019. 4.1

Brockwell, P., R. Davis. 1991. *Time series: theory and methods*. Springer. 2.3.1

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association* **100** 36–50. 1.1, 2.5.3, 3.1.1, 3.1.3, 3.1.4, 3.2, 3.2.1

Brown, L., L. hao. 2002. A test for the Poisson distribution. *Sankhya: The Indian Journal of Statistics* **64** 611–625. 3.4.1

Buzen, J. 1976. Fundamental operational laws of computer system performance. *Acta Informatica* **7**(2) 167–182. 2.1.1

Cady, N., M. Mattes, S. Burton. 1995. Reducing intensive care unit length of stay: A stepdown unit for first-day heart surgery patients. *Journal of Nursing Administration* **25**(12) 29–30. 4.2

Cameron, A., P. Trivedi. 1986. Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* **1**(1) 29–53. 4.4.3

Cameron, A., P. Trivedi. 1998. *Regression analysis of count data*. Cambridge University Press. 4.4.3

Chalfin, D., S. Trzeciak, A. Likourezos, B. Baumann, R. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35**(6) 1477–1483. 4.2

Chen, L., Edward H. Kennedy, A. Sales, T. Hofer. 2013. Use of health IT for higher-value critical care. *New England Journal of Medicine* **368** 594–597. 1, 1.2, 4.1, 4.2, 4.5.2

Chen, M., M. Render, A. Sales, E. Kennedy, W. Wiitala, T. Hofer. 2012. Intensive care unit admitting patterns in the veterans affairs health care system. *Archives of Internal Medicine* **172**(16) 1220–1226. 4.1

Cooper, R. 1982. *Introduction to Queueing Theory*. 2nd ed. North Holland, Amsterdam. 3.1

Daley, D., D. Vere-Jones. 2008. *An Introduction to the Theory of Point Processes*, vol. II. 2nd ed. Springer, Oxford, U. K. 3.1

Deb, P., P. Trivedi. 2006. Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment. *Stata Journal* **6**(2) 246–255. 4.4.3

Denning, P., J. Buzen. 1978. The operational analysis of queueing network models. *ACM Computing Surveys* **10**(3) 225–261. 2.1.1, 2.1.2, 2.2.3, 2.2.3

Durbin, J. 1961. Some methods for constructing exact tests. *Biometrika* **48**(1) 41–55. 3.1.3

Durbin Jr, C., R. Kopel. 1993. A case-control study of patients readmitted to the intensive care unit. *Critical Care Medicine* **21**(10) 1547–1553. 3

Eick, S., W. Massey, W. Whitt. 1993a. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* **39** 241–252. 2.3.2

Eick, S., W. Massey, W. Whitt. 1993b. The physics of the mt/g/ queue. *Operations Research* **41**(4) 731–742. 2.5.3

El-Taha, M., S. Stidham Jr. 1999. *Sample-Path Analysis of Queueing Systems*. Kluwer, Boston. 2.1

Escher, M., T. Perneger, J. Chevrolet. 2004. National questionnaire survey on what influences doctors' decisions about admission to intensive care. *BMJ* **329**(7463) 425–429. 4.2

Escobar, G., J. Greene, M. Gardner, G. Marelich, B. Quick, P. Kipnis. 2011. Intra-hospital transfers to a higher level of care: Contribution to total hospital and intensive care unit (icu) mortality and length of stay (los). *Journal of Hospital Medicine* **6**(2) 74–80. 4.3.2

Escobar, G., J. Greene, P. Scheirer, M. Gardner, D. Draper, P. Kipnis. 2008. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care* **46**(3) 232–239. 4.2, 4.1, 4.4.1

Feldman, Z., A. Mandelbaum, W. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54**(2) 324–338. 4.7

Fisher, E., D. Wennberg, T. Stukel, D. Gottlieb. 2004. Variations In The Longitudinal Efficiency Of Academic Medical Centers. *Health Affairs* . 4.2

Fralix, B., G. Riano. 2010. A new look at transient versions of Little's law. *Journal of Applied Probability* **47** 459–473. 1.1

Franklin, C., E. Rackow, B. Mamdani, G. Burke, M. Weil. 1990. Triage considerations in medical intensive care. *Archives of Internal Medicine* **150**(7) 1455. 4.2

Gans, N., G. Koole, A. Mandelbaum. 2003. Commissioned paper: Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141. 1

Glasserman, P., D. Yao. 1994. Monotone optimal control of permutable gsmps. *Mathematics of Operations Research* **19**(2) 449–476. 4.2

Glynn, P.., W. Whitt. 1986. A central-limit-theorem version of $l = \lambda w$. *Queueing Systems* **1**(2) 191–215. 2.2.2, 2.4.2

Glynn, P.., W. Whitt. 1987. Sufficient conditions for functional-limit-theorem versions of $l = \lambda w$. *Queueing Systems* **1**(3) 279–287. 2.4.2

Glynn, P., W. Whitt. 1989. Indirect estimation via $L = \lambda W$. *Operations Research* **37** 82–103. 2.4.2

Goldberg, D., W. Whitt. 2008. The last departure time from an $m_t/g/\infty$ queue with a terminating arrival process. *Queueing Systems* **58**(2) 77–104. 2.5.3

Green, L. 2003. How many hospital beds? *Inquiry* **39** 400–412. 4.1

Green, L., P. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16** 13–29. 2.6.2

Green, L., S. Savin, N. Savva. 2013. Nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Science, published online.* doi:10.1287/mnsc.2013.1713. 4.2

Halpern, L., N.and Bettes, R. Greenstein. 1994. Federal and nationwide intensive care units and healthcare costs: 1986–1992. *Critical Care Medicine* **22** 2001–2007. 4.1

Halpern, N., S. Pastores, H. Thaler, R. Greenstein. 2007. Critical care medicine use and cost among medicare beneficiaries 1995-2000: Major discrepancies between two united states federal medicare databases*. *Critical Care Medicine* **35**(3) 692–699. 4.1

Iapichino, G., D. Corbella, C. Minelli, G.H. Mills, A. Artigas, D.L. Edbooke, A. Pezzi, J. Kesecioglu, N. Patroniti, M. Baras, et al. 2010. Reasons for refusal of admission to intensive care and impact on mortality. *Intensive Care Medicine* **36**(10) 1772–1779. 4.1, 4.2

Ibrahim, R., P. L'Ecuyer, N. Regnard, H. Shen. 2012. On the modeling and forecasting of call center arrivals. *Proceedings of the 2012 Winter Simulation Conference* **2012** 256–267. 3.1.2, 3.1.4

Iezzoni, L., et al. 2003. *Risk adjustment for measuring health care outcomes*, vol. 3. Health Administration Press, Ann Arbor. 4.3.2

Jaeker, J.B., A.L. Tucker. 2013. An empirical study of the spillover effects of workload on patient length of stay. *Working Paper, Harvard Business School*. 4.2, 5

Jennings, O., A. Mandelbaum, W. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* **42** 1383–1394. 2.3.2

Jewell, W. 1967. A simple proof of $l = \lambda w$. *Operations Research* **15**(6) 1109–1116. 2.2.2

Johnson, N. 1978. Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association* **73**(363) 536–544. 2.6.1

Joint Position Statement. 1994. Essential provisions for critical care in health system reform. *Critical Care Medicine* **22** 2017–2019. 4.1

Jongbloed, G., G. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* **17** 307–318. 3.1.2

Kaplan, R., M. Porter. 2011. How to solve the cost crisis in health care. *Harvard Business Review* **89**(9) 10. 1, 4.1, 4.2

Kathirgamatamby, N. 1953. Note on the Poisson index of dispersion. *Biometrika* **40**(1) 225–228. 3.4.1

Kc, D., B. Staats. 2012. Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing and Service Operations Management* **14**(4) 618–633. 4.2

Kc, D., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498. 4.2, 4.3.2, 4.4.2, 4.5.2

Kc, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing and Service Operations Management* **14**(1) 50–65. 4.1, 4.2, 4.4.2, 4.4.2, 4.5.2, 4.5.2

Kim, S.-H., C. Chan, M. Olivares, G. Escobar. 2014. ICU admission control: An empirical study of capacity allocation and its implication on patient outcomes. *Working paper, Columbia University*. 1, 4

Kim, S.-H., W. Whitt. 2012a. Statistical analysis with Little's law: E-companion. 2.1.3, 2.3.1, 2.3.2, 2.3.3, 2.6.1

Kim, S.-H., W. Whitt. 2012b. Statistical analysis with Little's law: Technical report. 2.1.3, 2.3.3, 2.3.3, 2.3.4, 2.6.2, 2.6.3

Kim, S.-H., W. Whitt. 2013a. Appendix to are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? 3.1.4, 3.2.2, 3.3.4, 3.3.6, 3.4.1, 3.5.4

Kim, S.-H., W. Whitt. 2013b. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes?: Online supplement. 3.1, 3.1.3, 3.1.4, 3.2.2, 3.3.7

Kim, S.-H., W. Whitt. 2013c. Estimating waiting times with the time-varying Littles law. *Probability in the Engineering and Informational Sciences* **27** 471–506. 1.1, 2.1.3

Kim, S.-H., W. Whitt. 2013d. The power of alternative Kolmogorov-Smirnov tests based on transformations of the data. *Working Paper, Columbia University*. 1.1

Kim, S.-H., W. Whitt. 2013e. Statistical analysis with Little's law. *Operations Research* **61**(4) 1030–1045. 1, 1.1, 2, 3.3.1, 3.3.1, 3.5

Kim, S.-H., W. Whitt. 2013f. Statistical analysis with Little's law supplementary material: more on call center data. 3.3.1, 3.3.1, 3.5

Kim, S.-H., W. Whitt. 2014a. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Forthcoming in Manufacturing and Service Operations Management. 1, 1.1, 3

Kim, S.-H., W. Whitt. 2014b. Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics,* **61**(1) 66–90. 1, 1.1, 3.1.3, 3.1.4, 3.2.1, 3.2.1, 3.2.2, 3.3.1, 3.3.5, 3.4, 3.7

Kuntz, L., R. Mennicken, S. Scholtes. 2014. Stress on the ward: Evidence of safety tipping points in hospitals. *Forthcoming in Management Science*. 4.2

Lewis, P. 1965. Some results on tests for Poisson processes. *Biometrika* **52**(1) 67–77. 1.1, 3.1.3

Li, A., W. Whitt. 2013. Approximate blocking probabilities for loss models with independence and distribution assumptions relaxed. *Performance Evaluation,* Available online, August 13, 2013. 3.1

Little, J. 1961. A proof of the queueing formula: $L = \lambda W$. *Operations Research* **9** 383–387. 2.1, 2.1.1

Little, J. 2011. Little's law as viewed on its 50^{th} anniversary. *Operations Research.* **59** 536–539. 2.1, 2.1.1, 2.1.2, 2.2.3, 2.2.3

Little, J., S. Graves. 2008. Little's law. D. Chhajed, T. J. Lowe, eds., *Building Intuition: Insights from Basic Opertations Management Models and Principles*, chap. 5. New York, 81–100. 2.1.2

Litvak, N., M. van Rijsbergen, R. Boucherie, M. van Houdenhoven. 2008. Managing the overflow of intensive care patients. *European Journal of Operational Research* **185** 998–1010. 3.1

Liu, V., P. Kipnis, M. Gould, G. Escobar. 2010. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Medical care* **48**(8) 739. 4.4.1

Louriz, M., K. Abidi, M. Akkaoui, N. Madani, K. Chater, J. Belayachi, T. Dendane, A.A. Zeggwagh, R. Abouqal. 2012. Determinants and outcomes associated with decisions to deny or to delay intensive care unit admission in morocco. *Intensive Care Medicine* 1–8. 4.1, 4.2

Lovejoy, W., J. Desmond. 2011. Little's law flow analysis of observation unit impact and sizing. *Academic Emergency Medicine* **18** 183–189. 2.1.2

Luyt, C., A. Combes, P. Aegerter, B. Guidet, J. Trouillet, C. Gibert, J. Chastre. 2007. Mortality among patients admitted to intensive care units during weekday day shifts compared with off hours. *Critical Care Medicine* **35**(1) 3–11. 4.3.2

Mandelbaum, A. 2010. Lecture notes on Little's law, course on service engineering. The Technion, Israel, http://iew3.technion.ac.il/serveng/Lectures/lectures.html. 2.1.2, 2.2.2

Mandelbaum, A. 2012. Service Engineering of Stochastic Networks web page: http://iew3.technion.ac.il/serveng/. 2.3, 3.5

Marsaglia, G., W. Tsang, J. Wang. 2003. Evaluating Kolmogorov's distribution. *Journal of Statistical Software* **8**(18) 1–4. 3.3.2

Massey, F. 1951. The Kolmgorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* **461** 68–78. 3.3.2

Massey, W., G. Parker, W. Whitt. 1996. Estimating the parameters of a nonhomogeneous Poisson process with linear rate. *Telecommunication Systems* **5** 361–388. 3.1.4, 3.3

Miller, B. 1969. A queueing reward system with several customer classes. *Management Science* **16**(3) 234–245. 4.2

Miller, L. 1956. Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association* **51** 111–121. 3.3.2

Mullan, F. 2004. Wrestling With Variation: An Interview With Jack Wennberg. *Health Affairs* **Suppl Variation** VAR73–80. 4.1, 4.2

O'Connor, A., H. Llewellyn-Thomas, A. Flood. 2004. Modifying Unwarranted Variations In Health Care: Shared Decision Making Using Patient Decision Aids. *Health Affairs* **Suppl Variation** VAR63–72. 4.2

OECD. 2014. Health expenditure. URL http://www.oecd.org/els/health-systems/health-expenditure.htm. 1

Osadchiy, N., V. Gaur, S. Seshadri. 2013. Sales forecasting with financial indicators and experts' input. *Production and Operations Management* **22**(5) 1056–1076. 4.2

Pang, G., W. Whitt. 2012. The impact of dependent service times on large-scale service systems. *Manufacturing and Service Operations Management* **14**(2) 262–278. 3.1

Papastavrou, J., S. Rajagopalan, A. Kleywegt. 1996. Discrete dynamic programming and capital allocation. *Management Science* **42** 1706–1718. 4.2

Phillips, R., A. Simsek, G. Van Ryzin. 2013. Does field price discretion improve profits? evidence from auto lending. *Working Paper, Columbia Business School*. 4.2

Pronovost, P., D. Needham, H. Waters, C. Birkmeyer, J. Calinawan, J. Birkmeyer, T. Dorman. 2004. Intensive care unit physician staffing: Financial modeling of the leapfrog standard*. *Critical Care Medicine* **32**(6) 1247–1253. 4.1

Raftery, A. 1995. Bayesian model selection in social research. *Sociological methodology* **25** 111–164. 4.6.3

Ramdas, K., K. Saleh, S. Stern, H. Liu. 2012. New joints more hip? learning in the use of new components. *Working Paper, London Business School*. 4.2

Reignier, J., Romain, S. Katsahian, L. Martin-Lefevre, B. Renard, M. Fiancette, C. Lebert, E. Clementi, F. Bontemps. 2008. Patient-related factors and circumstances surrounding decisions to forego life-sustaining treatment, including intensive care unit admission refusal*. *Critical care medicine* **36**(7) 2076–2083. 7

Reis Miranda, D., M. Jegers. 2012. Monitoring costs in the ICU: a search for a pertinent methodology. *Acta Anaesthesiologica Scandinavica* **56**(9) 1104–13. 4.1

Robert, R., J. Reignier, C. Tournoux-Facon, T. Boulain, O. Lesieur, V. Gissot, V. Souday, M. Hamrouni, C. Chapon, J. Gouello. 2012. Refusal of intensive care unit admission due to a full unit impact on mortality. *American journal of respiratory and critical care medicine* **185**(10) 1081–1087. 4.2

Serfozo, R. 1999. *Introduction to stochastic networks*, vol. 44. Springer. 2.1

Shi, P., M. Chou, J. Dai, D. Ding, J. Sim. 2012. Hospital inpatient operations: Mathematical models and managerial insights. *Working paper, Georgia Institute of Technology*. 1, 4.3.2

Shmueli, A., M. Baras, C. Sprung. 2004. The effect of intensive care on in-hospital survival. *Health Services and Outcomes Research Methodology* **5**(3) 163–174. 4.1, 4.2, 4.4.2

Shmueli, A., C. Sprung. 2005. Assessing the in-hospital survival benefits of intensive care. *International Journal of Technology Assessment in Health Care* **21**(01) 66–72. 4.2

Shmueli, A., C. Sprung, E. Kaplan. 2003. Optimizing admissions to an intensive care unit. *Health Care Management Science* **6**(3) 131–136. 4.2, 4.6.1

Sigman, K. 1995. *Stationary marked point processes: an intuitive approach*, vol. 134. Chapman & Hall New York. 2.1

Simard, R., P. L'Ecuyer. 2011. Computing the two-sided Kolmogorov-Smirnov distribution. *Journal of Statistical Software* **39**(11) 1–18. 3.1.1, 3.3.2, 2, 3.3.7

Simchen, E., C. Sprung, N. Galai, Y. Zitser-Gurevich, Y. Bar-Lavi, G. Gurman, M. Klein, A. Lev, L. Levi,

F. Zveibil, et al. 2004. Survival of critically ill patients hospitalized in and out of intensive care units under paucity of intensive care unit beds. *Critical Care Medicine* **32**(8) 1654–1661. 4.1, 4.2

Simpson, H., M. Clancy, C. Goldfrad, K. Rowan. 2005. Admissions to intensive care units from emergency departments: a descriptive study. *Emergency Medicine Journal* **22**(6) 423–428. 4.1

Singer, D., P. Carr, A. Mulley, G. Thibault. 1983. Rationing intensive care physician responses to a resource shortage. *New England Journal of Medicine* **309**(19) 1155–1160. 4.2

Sprung, C., D. Geber, L. Eidelman, M. Baras, R. Pizov, A. Nimrod, A. Oppenheim, L. Epstein, S. Cotev. 1999. Evaluation of triage decisions for intensive care admission. *Critical Care Medicine* **27**(6) 1073. 4.1

Stidham, S., Jr. 1974. A last word on $L = \lambda W$. *Operations Research* **22** 417–421. 2.1

Strand, K., H. Flaatten. 2008. Severity scoring in the icu: a review. *Acta Anaesthesiologica Scandinavica* **52**(4) 467–478. 4.2

Strauss, M., J. LoGerfo, J. Yeltatzie, N. Temkin, L. Hudson. 1986. Rationing of intensive care unit services. *Journal of the American Medical Association* **255**(9) 1143–1146. 4.2

Tafazzoli, A., N. Steiger, J. Wilson. 2011. N-skart: A nonsequential skewness- and autoregression-adjusted batch-means procedure for simulation analysis. *IEEE Transactions on Automatic Control* **56**(2) 254–264. 2.1.2, 2.4, 2.6.1

Tafazzoli, A., J. Wilson. 2010. Skart: A skewness-and autoregression-adjusted batch-means procedure for simulation analysis. *IIE Transactions* **43**(2) 110–128. 2.1.2, 2.4, 2.6.1

Task Force of the American College of Critical Care Medicine, Society of Critical Care Medicine. 1999. Guidelines for intensive care unit admission, discharge, and triage. *Critical Care Medicine* **27** 633–638. 4.1

The Kaiser Family Foundation, statehealthfacts.org. 2012. Hospital adjusted expenses per inpatient day, 2009. URL http://www.statehealthfacts.org/. 4.6.4

Van Walraven, C., G. Escobar, J. Greene, A. Forster. 2010. The kaiser permanente inpatient risk adjustment

methodology was valid in an external patient population. *Journal of Clinical Epidemiology* **63**(7) 798–803. 4.2

Vanhecke, T., M. Gandhi, P. McCullough, M. Lazar, K. Ravikrishnan, P. Kadaj, R. Begle. 2008. Outcomes of patients considered for, but not admitted to, the intensive care unit*. *Critical Care Medicine* **36**(3) 812–817. 4.2

Veatch, M., L. Wein. 1992. Monotone control of queueing networks. *Queueing Systems* **12**(3-4) 391–408. 4.2

Weber, R., S. Stidham Jr. 1987. Optimal control of service rates in networks of queues. *Advances in Applied Probability* **19**(1) 202–218. 4.2

Weinstein, J., K. Bronner, T. Morgan, J. Wennberg. 2004. Variations in Major Surgery for Degenerative Diseases of the Hip, Knee, and Spine. *Health Affairs* **Suppl Variation** VAR81–89. 4.2

Wennberg, J., E. Fisher, J. Skinner, et al. 2002. Geography and the debate over medicare reform. *Health Affairs* **21**(2) 10–10. 6

Whitt, W. 1982. Approximating a point process by a renewal process: two basic methods. *Operations Research* **30** 125–147. 3.2.2

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York. 3.1

Whitt, W. 2012. Extending the FCLT version of $L = \lambda W$. *Operations Research Letters* **40**(4) 230–234. 2.4.2

Wilkinson, R. 1956. Theories of toll traffic engineering in the U.S.A. *Bell System Technical Journal* **35** 421–514. 3.1

Willink, R. 2005. A confidence interval and test for the mean of an asymmetric distribution. *Communications in Statistics Theory and Methods* **34**(4) 753–766. 2.6.1

Wooldridge, J. 2010. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, MA. 4.4.3, 4.5.2

Yom-Tov, G., A. Mandelbaum. 2014. The Erlang-R queue: time-varying QED queues with re-entrant

customers in support of healthcare staffing. *Forthcoming in Manufacturing and Service Operations Management* . 4.7

Ziser, A., M. Alkobi, R. Markovits, B. Rozenberg. 2002. The postanaesthesia care unit as a temporary admission location due to intensive care and ward overflow. *British Journal of Anaesthesia* **88**(4) 577–579. 4.2