PLANNING AND SCHEDULING SURGERIES UNDER STOCHASTIC ENVIRONMENT

A Dissertation

by

SANGDO CHOI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Wilbert E. Wilhelm
Committee Members,	Amarnath Banerjee
	Lewis Ntaimo
	Kiavish Kianfar
	Michael Ketzenberg
Department Head,	Cesar Malave

December 2012

Major Subject: Industrial Engineering

ABSTRACT

This dissertation presents an integrated approach to planning and scheduling surgeries in operating-rooms (ORs) at strategic, tactical and operational levels. We deal with uncertainties of surgery demand and durations to reflect a reality in OR management.

The strategic part of the dissertation studies capacity decisions that allocate surgical specialties to OR days with the objective of minimizing total expected costs due to penalties for any patients who are not accommodated and for under- (i.e., idleness) and over- (i.e., overtime) usage of OR capacity. It presents a prototypical non-linear, stochastic programming model to structure the problem and four adaptations, along with associated solution approaches, with the goal of facilitating solution by overcoming the computational disadvantages of the prototype. Each of these models offers advantages but is also attended by disadvantages. Computational tests compare the four models and solution approaches with respect to solution quality and run time.

The tactical part of the dissertation prescribes an approach to optimize a master surgical schedule (MSS), which adheres to the block scheduling policy, using a new type of newsvendor-based model. Our newsvendor approach prescribes the optimal duration of each block and the best permutation, obtained by solving the *sequential newsvendor* problem, determines the optimal block sequence. We obtain closed-form solutions for the case in which surgery durations follow the normal distribution. Furthermore, we give a closed-form solution for optimal block duration with no-shows. We conduct numerical tests for surgery durations that follow normal, lognormal and gamma distributions. Results show that the closed-form solutions associated with the normal distribution gives close approximations to solutions associated with log-

normal and gamma distributions.

The operational part of the dissertation prescribes an optimal rule to sequence two or three surgeries in a block. The smallest-variance-first-rule (SV) is generally accepted as the optimal policy for sequencing two surgeries, although it has been proven formally only for several restricted cases. We extend prior work, studying three distributions as models of surgery duration (the lognormal, gamma, and normal) and including overtime in a total-cost objective function comprising surgeon-and-patient-waiting-, operating-room-idle-, and staff over-times. We specify expected waiting- and idle- time as functions of the parameters of surgery duration to identify the best rule to sequence two surgeries. We compare the relative values of expected waiting- and idle- times numerically with that of expected overtime. Results recommend that the SV rule be used to minimize total expected cost of waiting-, idle- and over-time. We find that gamma and normal distributions with the same mean and variance as the lognormal give nearly the same expected waiting- and idle- times, observing that the lognormal in combination with either the gamma or normal gives a similar result.

Lastly, the dissertation investigates an appointment system with deterministic arrival times (D) and non-identical exponential service times ($\tilde{M}$). For two customers, we show that both the smallest-mean-first-rule and the SV minimize the sum of expected waiting- and idle-times. We prove that neither is optimal for three customers, but verifies that the first customer in the sequence should be the one with the smallest variance (mean).

DEDICATION

Dedicated to my precious daughters, Hannah, Deborah, Joanne and beloved wife, Yoony.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Professor Wilbert E. Wilhelm for his continuous guidance and teaching of how to write a paper. Through the editing process, much of the wording in this dissertation was edited and composed by Professor Wilhelm. Without his support I could not have completed my Ph.D. studies. Various chapters of this dissertation have been or will be published in refereed journals, which will hold the associated copyrights.

I owe a special acknowledgment to Dr. Amarnath Banerjee, Dr. Lewis Ntaimo, Dr. Kiavash Kianfar, and Dr. Michael Ketzenberg for serving on the committee and for giving helpful suggestions. All of my committee members have stimulated me intellectually in their research areas, and I have expanded my research areas purely due to the working experience with them.

I would like to thank Dr. Elym Tekin, who as an ex-advisor, had supported my doctoral studies financially for two years. I would also like to thank to Dr. Curry and our Department for allowing me to teach for six semesters, which gave me the financial freedom to conduct my work. I am especially grateful to Judy Meeks for always being so kind in assisting me in many different ways.

Most importantly, I would like to thank my family for their patience and sacrifice. They are the greatest source of pleasure and relaxation in my life ever.

NOMENCLATURE

CA	Capacity Allocation
CLT	Central Limit Theorem
CPT	Current Procedure Terminology
CV	Coefficient of Variation
ENT	Ear Nose Throat
KKT	Karush Kuhn Turker
LM	Largest Variance First Rule
MSS	Master Surgical Schedule
NL-CA	Non-Linear Capacity Allocation
NS-SIP	Non Symmetric Stochastic Integer Programming
NV-CA	News Vendor Capacity Allocation
NV-SIP	News Vendor Stochastic Integer Programming
OR	Operating Room
OTP	Overtime Penalty
PACU	Post Anesthesia Care Unit
SICU	Surgical Intensive Care Unit
SIP	Stochastic Integer Programming
SM	Smallest Mean First Rule
SNV	Sequential News Vendor
SV	Smallest Variance First Rule
SWIP	Sum of Waiting- and Idle-Time Penalties
USNV	Unconstrained Sequential News Vendor

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

This dissertation prescribes an integrated way for planning and scheduling surgeries in operating room (OR). We deal with capacity allocation decision at the strategic level; master surgical block schedule at the tactical level; sequencing surgeries at the operational level. The decisions at a higher level are used as a constraint to lower level decisions. For example, tactical decision determines the planned block duration of a sub-specialty and operation decision determines sequencing surgeries of the assigned sub-specialty. We also analyze operational decisions under a general appointment scheduling system.

Each OR provides vital services to patients and a major source of revenue to the hospital; it employs capital-intensive equipment and skilled surgery teams (e.g., surgeons, anesthesiologists, nurses), who are highly paid. Hospital administrators seek to utilize capacities (e.g., capital-intensive equipment and human resources) as efficiently as possible.

Every hospital provides a unique capacity for performing surgery through the numbers of ORs and surgical skills it offers. A surgical suite typically comprises several ORs, each of which is equipped to support one (e.g., heart, neurological, or orthopedic) or several (e.g., general surgery, ENT) specialties. The typical surgical specialty comprises a number of sub-specialties. For instance, the orthopedics specialty includes hip replacement, knee replacement, femur fixation, and shoulder repair sub-specialties. Surgeries that require the same sub-specialty are medically homogeneous and require the same medical expertise of the surgeon or perhaps a group of surgeons who practice the same sub-specialty (van Oostrum et al., 2008).

OR capacity is typically measured by three components: physical resources, hu-

man resources, and time availability (May et al., 2011). Physical resources include the number of ORs and the equipment installed in each OR. Some surgical specialties (e.g., cardiology, neurology, orthopedics) require specialized equipment that, when installed in an OR, dedicates that OR to that particular specialty. Some specialties (e.g., general surgery, ENT) require less specialized equipment and can share ORs that provide such flexibility. Human resources, which include surgeons, anesthesiologists, and nurses, can be assigned to ORs as desired. Time availability at the strategic level relates to the length of the OR work day (e.g., 8 hours), and, at the tactical level, to time blocks, which are shorter (e.g., 2 or 4 hours). We concentrate on time availability to manage OR capacity, assuming that physical and human resources are fixed.

Capacity planning is a process of specifying the levels of resources necessary to meet demands in a cost-effective way (Blake, 2011). Inadequate capacity planning can deteriorate the quality of care provided by hospitals (Bai et al., 2009). For example, hospital administrators may have to meter patient admissions over time or route patients to other hospitals if capacity is not sufficient to accommodate them. Capacity planning over different time horizons involves constructing and/or upgrading facilities (very long term, or strategic), allocating specialties to OR days (long term, or strategic), assigning sub-specialties to time blocks (medium term, or tactical), scheduling actual patients within time blocks in a specific OR day (short term, or operational), making last minute adjustments (very short term, or real-time), and executing the schedule (contemporaneous) (May et al., 2011). We focus on the long term decisions that allocate specialties to OR days, not capacity expansion.

1.1 Research Scope

Strategic-level allocations provide a structure within which tactical master surgical schedules (MSSs) are prescribed (Choi and Wilhelm, 2012b) to assign subspecialties to time blocks in each OR day to which the associated specialty is allocated. Should the intermediate-term forecast prepared to plan the MSS differ substantially from the long-range forecast upon which strategic allocations have been based, the allocation model can be implemented again using the refined, tactical forecast so that the MSS is consistent with specialty-to-OR-day allocations. Subsequently, operational level decisions schedule specific patients within the time blocks prescribed by the MSS.

Strategic level deals with specialty in a horizon of long term; tactical, subspecialty, medium term; operational, individual surgery, short term. Table 1.1 illustrates exemplar inputs and main decisions by a decision hierarchy based on decision level and time frame, accordingly. Capacity expansion, bed planning, assignment of subspecialties, time tabling, rescheduling, and execution are out-of-scope, in *italics* in Table 1.1. While capacity expansion is an ad-hoc decision, we deal with capacity planning of on-going basis. Bed planning decision is more related to general capacity decision rather than OR capacity decision. If the allocation decision is made at the strategic level, it is not necessary to assign sub-specialty to ORs or blocks at the tactical level. We focus on only sequencing decision other than scheduling decisions such as time tabling or rescheduling.

We assume an environment in which strategic decisions assign specialties to OR days based on a long-term (e.g., annual) forecast. MSS decisions assign subspecialties to time blocks within each day, based on an intermediate-term forecast, which can be expected to be more accurate because it deals with a shorter time hori-

Table 1.1: Decision Hierarchy from Long-term to Short-term

Decision Level	Time Frame	Inputs	Main Decisions
Strategic (specialty)	Long term	Demand forecast Surgery duration Standing schedule Contribution margin Related costs	Capacity allocation Patient mix *OR times estimation* *Capacity expansion* *Bed planning*
Tactical (subspecialty)	Medium term	Allocation of specialties Demand forecast/actual lists Surgery duration Related costs	Block duration and sequence *Assignment of subspecialties*
Operational (surgery)	Short term	Block duration and start time Surgery duration Related costs	Sequencing surgeries *Time tabling* *Rescheduling* *Execution*

zon and may include a mix of actual and forecast needs. Operational-level decisions, which are made on a daily basis, assign actual patients to specific times within time blocks, matching sub-specialty need with the MSS (e.g. Dexter et al. (2005)). We depict our focus of the dissertation as shown Figure 1.1.

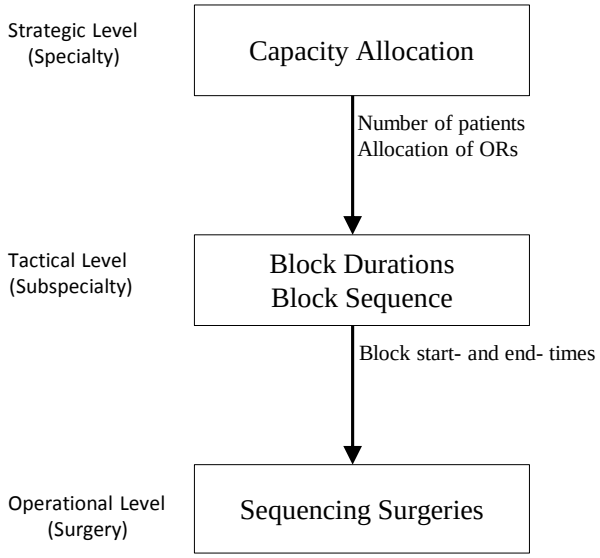


Figure 1.1: Main Decision Variables by Level

The following subsections describe the problem background of individual parts: strategic, tactical, and operational, respectively.

1.1.1 Capacity Allocation at the Strategic Level

We begin by formulating a prototypical non-linear, stochastic model to identify relevant practical features of the problem and to structure them. The prototype incorporates nonlinear forms of several types and is not computationally attractive. Thus, we propose four adaptations to linearize it with the goal of facilitating solution: NV-CA, NV-SIP, SIP and NS-SIP. The first two (i.e., NV-CA, NV-SIP) are based on the *inverse news vendor* model (Carr and Lovejoy, 2000), each of the last three (i.e., NV-SIP, SIP, and NS-SIP) recast decision variables and involve stochastic programs for which we adopt *recourse* models to prescribe certain decisions that must be delayed until stochastic processes are realized, resolving uncertainty. A recourse model is designed to prescribe a solution that balances the potential impacts of various possible outcomes (Higle, 2005).

1.1.2 Master Block Surgical Schedule at the Tactical Level

Tactical-level decisions for the intermediate-term (e.g., month or quarter) prescribe an MSS to assign sub-specialties to time blocks in each OR each day. Under a block scheduling policy, an MSS must determine block duration and sequence for each OR day, to minimize the total cost of earliness and tardiness.

We deal with the block scheduling policy in this study. A block is the amount of time during which an OR is assigned to a specific sub-specialty. For example, a block may be planned with the duration of two hours, half of a day (e.g., morning, or afternoon), or a day-long duration, for example, to permit a surgeon to perform a series of complex surgeries. An alternative, the open scheduling policy, under which each surgeon can schedule his/her surgeries at any time, was common in the 1960s

and 1970s but is rarely used in practice today, because it does not utilize surgeons' time as efficiently as block scheduling (Blake et al., 2002).

An MSS, which is analogous to a master production schedule in a manufacturing environment, has a number of important uses. MSS defines aggregate resource requirements of peri-operative activities and ancillary departments (e.g., post-anesthesia care unit(PACU), surgical intensive care unit(SICU), nursing), not only of ORs and surgeons. Nurse managers should ensure that the set of ORs and PACUs run compatibly each day of the week (Blake and Donald, 2002) so that actual decisions adhere to the MSS as strictly as possible. Like Dexter and Hopwood (1999), Rohleder et al. (2005), and Samanlioglu et al. (2010), this paper focuses on ORs and does not deal with other departments. MSS enables hospital managers to respond to random events (e.g., a short-term shortage of surgeons or anesthetists), seasonal fluctuations in demand (e.g., summer or Christmas time), or strategic decisions that alter program emphasis (e.g., to respond to an increasing popularity of cosmetic surgery) (Blake and Donald, 2002). In particular, the operational-level uses the MSS to schedule individual patients; if actual demand levels were to deviate significantly from the MSS, a hospital manager should update the MSS to better accommodate them.

1.1.3 Sequencing Surgeries in a Block at the Operational Level

Sequencing surgeries in a single OR involves only a few patients in each time block, which may, for example, have a duration of two-, four-, or eight-hours. Each surgery requires a random duration that depends upon its type (i.e., specialty such as cardiac, orthopedic, or neurological) of surgery. In order to study sequencing policies for different surgery-duration distributions, the dissertation deals with two or three surgeries in a time block of duration h and analyzes three different distributions

of surgery duration (lognormal, gamma, and normal), rather than invoking general restrictions such as stochastic order or increasing (decreasing) hazard rate, as have previous studies (Niño-Mora, 2002; Gupta, 2007; Denton et al., 2007; Pinedo, 2009).

We note that most previous studies assumed that surgery durations are independent and identically distributed (*i.i.d.*) for all patients (Cayirli and Veral, 2003). We assume that surgery durations are independent, but extend prior results, allowing durations that are not identically distributed. We focus on sequencing surgeries that require the same specialty for a given set of patients in a single time block.

1.2 Research Objectives of the Dissertation

The primary objective of the dissertation is an integrated way of planning and scheduling surgeries at strategic, tactical, and operational level. Each part in the dissertation achieves its own research objectives for each level. We describe research objectives at the strategic, tactical, and operational levels, respectively.

The research objectives of strategic part are (1) a prototypical model to optimize the allocation of surgical specialties to OR days, resulting in the mix of patients accommodated; (2) four adaptations to facilitate, along with associated solution approaches: news vendor-based capacity allocation (NV-CA), news vendor-based stochastic integer programming (NV-SIP), stochastic integer programming (SIP) and stochastic integer programming without symmetry (NS-SIP); and (3) numerical tests that compare the computational characteristics of the four models.

Tactical-level decisions for the intermediate-term (e.g., month or quarter) prescribe an MSS, which assigns surgery sub-specialties to time blocks in each operating room (OR) each day. Under a block scheduling policy, an MSS must determine block duration and sequence for each OR day to minimize the sum of expected earliness and tardiness costs. With the goal of synthesizing a methodology to prescribe an MSS,

specific research objectives of this paper are (1) a method to optimize the planned duration of each block, minimizing the sum of expected earliness and lateness costs; (2) a method to optimize the sequence (i.e., permutation) of blocks in each OR day; and (3) a method to prescribe an optimal planned block duration when no-shows are considered.

The research objectives of the operational part are (1) identifying the best rule to prescribe the sequence of two surgeries; (2) specifying expected waiting- and idle-time as functions of the parameters of surgery duration; (3) numerically comparing the relative values of expected waiting and idle times with that of expected overtime; (4) extending to the case in which a lognormal distribution is combined with either a gamma or a normal distribution; (5) modeling the three-surgery case with normally distributed durations; and (6) demonstrating how our results can be applied by using them as a basis for a heuristic that assigns surgeries to multiple ORs and sequences them in each OR.

1.3 Contributions of the Dissertation

The dissertation contributes from several perspectives. It shows how upper-level decisions affect (or constrain) lower-level decisions in an integrated way. Strategic and tactical levels deal with assignment problems of specialties and subspecialties, respectively. The latter determines blocks durations of an OR, given the former decisions. Block durations at the tactical level are more refined than the OR time at the strategic level. Operational-level decisions are constrained by both the number of assigned surgeries at the strategic-level decision and block durations at the tactical-level decision.

The dissertation prescribes two new types of newsvendor model: *inverse newsvendor* and *sequential newsvendor*. The former model of the strategic part determines

the optimal number of assigned surgeries, and can be applied to a setting in which one may assign customers (or orders) to a capacitated resource, such as airline booking. The latter model of the tactical part prescribes the optimal duration and sequences of blocks, and can be extended to a setting in which one may determine production (or delivery) times and sequence. *Sequential newsvendor* is a series of time-based newsvendor problems, not a quantity-based multi-period problem, in which quantities are random and each interval is of fixed duration.

Lastly, the dissertation conducts numerical studies to support or complement analytical results, which cover general instances with wide ranges of mean and variance. The strategic part shows the effectiveness of the proposed heuristic using a large number of scenarios. The tactical part prescribes block durations and sequence analytically using the normal distribution, and conducts numerical studies to show that the gamma or the lognormal has nearly the same results. The operational part conducts extensive numerical studies to convince analytical results.

1.4 Outline of the Dissertation

The remainder of the dissertation is organized as follow. Chapter 2 reviews strategic-level decisions, tactical-level decisions, and sequencing surgeries at the operational level, respectively. Chapter 3 prescribes four solving models for capacity allocation problem. Chapter 4 prescribes a master block surgical scheduling approach. Chapter 5 prescribes optimal rule to sequence surgeries in a block. Chapter 6 prescribes optimal rule to sequence customers in an appointment scheduling system in which the inter-arrival time is deterministic. Chapter 7 gives conclusions and offers suggestions for future research.

2. LITERATURE REVIEW

A review of the literature suggests few studies have addressed strategic OR capacity planning. The literature deals almost entirely with questions of medium- to short-term planning in which it is assumed that the number of ORs is fixed (Green, 2004).

Several studies (Blake and Carter, 1997; Gupta, 2007; Cardoen et al., 2010) have proposed three-level classifications of OR planning and scheduling: strategic (long-term), tactical (medium-term), and operational (short-term). May et al. (2011) updated previous classification schemes, expanding their scope to encompass a number of emerging topics such as scheduling (rescheduling) the day on which each surgery is to be performed. We focus on strategic decisions that allocate surgical specialties to OR days.

We review literatures at strategic-, tactical-, and operational-level, respectively in the following sections.

2.1 Strategic Level Decisions

A review of the literature suggests that few studies have addressed strategic OR capacity planning. The literature deals almost entirely with questions of medium- to short-term planning in which it is assumed that the number of ORs is fixed (Green, 2004).

A few authors (Dexter et al., 2005; Dexter and O'Neill, 2004; Bai et al., 2009) studied capacity expansion at the strategic level. Dexter et al. (2005) explored the allocation of OR time after the decision has been made to increase the number of ORs. Dexter and O'Neill (2004) applied data envelopment analysis (DEA) in several contexts pertaining to capacity expansion, workload, and external competition.

Bai et al. (2009) investigated the role of accounting and operational factors as well as interactions among these factors that drive OR capacity investments, combining insights from analytical models in the accounting and operations management literature. In contrast to these approaches, we focus on capacity allocation rather than capacity expansion.

Other studies have dealt with allocation as a tactical, medium-term problem (Wachtel and Dexter, 2008; Strum et al., 1997; Dexter et al., 2002; Kuo et al., 2003). Wachtel and Dexter (2008) noted that assigning surgical sub-specialties to expanded OR capacities is a tactical decision. Strum et al. (1997) formulated a news vendor model to determine OR utilization, analyzing the quality of surgical schedules, and allocating surgical budgets. Dexter et al. (2002) described the allocation of OR time from a financial perspective, and determined the mix of allocations that maximizes excess revenue. Kuo et al. (2003) used linear programming to allocate OR time among a group of surgeons based on the fees that would be generated. In contrast, we derive our allocation models for use at the strategic level with the presumption that resulting allocation decisions provide a structure in which a tactical MSS assigns sub-specialties to time blocks each OR day.

2.2 Tactical Level Decisions

Few studies have addressed MSS. Complicating matters, there is no commonly accepted standard definition of MSS (Testi et al., 2007; van Oostrum et al., 2008). Blake and Donald (2002), Santibanez et al. (2007), and Fei et al. (2008) described the MSS development process in detail, comparing it with master production scheduling in manufacturing. van Oostrum et al. (2008) discussed the pros and cons of MSS, compared centralized and decentralized MSS-planning processes, addressed various implementation issues and discussed suitability for hospitals with different organiza-

tional foci and culture.

Due to the absence of a standard definition, various studies have assigned surgeries to ORs as part of strategic, tactical, or operational decisions. The strategic problem of assigning specialties to ORs assumes that each OR day comprises a single time block and determines the number of OR-days for each specialty. One line of research on intermediate-term decisions has investigated assigning the expected number of surgeries associated with each specialty to OR days. In contrast, we regard this assignment problem as a strategic-level decision and assume that the assignment of specialties to ORs is given. Santibanez et al. (2007) assigned specialties to time blocks at the tactical-level, assuming that both the total amount of OR time and the number of patients are predetermined for each specialty over the planning horizon. Following Santibanez et al. (2007), our approach invokes the assumption that the number of patients is forecast for each sub-specialty. Guinet and Chaabane (2003) and Jebali et al. (2006) combined the assignment of specialties to ORs, typically a strategic-level problem, and the sequencing of surgeries in each OR, considered an operational-level issue, in one model.

A number of OR methodologies have been used to assign surgeries to ORs or blocks. Both deterministic integer programs (Kharraja et al., 2006; Blake and Donald, 2002; Zhang et al., 2009; Fei et al., 2009) and stochastic programs (Denton et al., 2010; Beliën et al., 2009) have been used to prescribe MSSs. Kharraja et al. (2006) modeled the assignment of specialties to days of pre-specified duration as a cutting stock problem with the objective of minimizing penalties for under- and over-use of ORs. Blake and Donald (2002) and Zhang et al. (2009) developed an analytical solution and incorporated it in a simulation model that captures randomness (e.g., random arrivals, no-shows) and non-linearities (e.g., non-proportional allocation of demand). Fei et al. (2008) studied surgery assignment using a set-partitioning for-

mulation and branch-and-price. Denton et al. (2010) and Beliën et al. (2009) used stochastic optimization at the operational level to assign surgeries to ORs on a given day.

A number of studies have used newsvendor models to prescribe block duration. Several studies (Strum et al., 2000a; Olivares et al., 2008; Wachtel and Dexter, 2010) have employed the newsvendor model to optimize the duration of a single block; they do not deal with sequencing blocks. This approach is more closely related to ours than is the assignment problem used, for example, by Guinet and Chaabane (2003) and Jebali et al. (2006). Guerriero and Guido (2010) and May et al. (2011) employed a newsvendor model at the strategic level to determine OR time for a specialty. Strum et al. (2000a) developed a newsvendor model to find the optimal block duration based on historical workloads (e.g., numbers of surgeries performed, numbers of staff hours). Olivares et al. (2008) applied a newsvendor model to determine how much OR time to reserve for a specific cardiac surgery to balance the costs of reserving too much vs too little OR time. Wachtel and Dexter (2010) gave a systematic review of the behavioral and experimental literature associated with newsvendor problems relevant to OR management and commented on the potential significance of these studies to OR management.

In contrast to earlier studies, we employ a newsvendor model to prescribe planned end-time (accordingly, block durations as well) and the *sequential newsvendor* model to specify block sequence. No prior research has addressed the block-sequence problem. We are the first to provide a closed-form solution for the case in which surgery durations are independent and normally distributed. Using the closed form that we obtain, we are able to derive the optimal rule to sequence blocks.

2.3 Operational Level Decisions

Researchers in stochastic scheduling typically seek to optimize an overall measure of schedule performance such as the sum of expected completion times or expected makespan. In contrast, OR scheduling focuses on minimizing waiting-, idle- and over-time penalties. Niño-Mora (2002) and Pinedo (2008) summarized stochastic scheduling research. Righter (1994) provided a review of stochastic ordering and its application in scheduling. One fundamental result for the single-machine configuration has shown that the rule that schedules the job with the smallest-mean-first-rule (SM) minimizes the sum of completion times under the assumption that all job processing times are independent and exponentially distributed (Glazebrook, 1979); that they have a common, general distribution with a nondecreasing hazard rate function (Weber, 1982); or that they follow stochastically ordered distributions (Weber et al., 1986). The largest-mean-first-rule (LM) rule minimizes expected makespan for the single-machine configuration when job processing times are exponentially distributed (Bruno et al., 1981), or when job processing times follow a common distribution with a nondecreasing hazard rate function (Weber, 1982). Although many articles on stochastic scheduling impose strict assumptions, including, for example, that service time is exponentially distributed, we consider three distributions that are relevant to surgery scheduling: lognormal, gamma, and normal.

Appointment-based models (e.g., health care, law firm) typically assume that customers arrive for service at pre-determined, rather than random, times (Wang, 1993). Gupta and Denton (2008) summarized key issues in appointment systems for health services. Jansson (1966) studied the D/M/1 queueing model of appointment systems. Wang (1993) and Wang (1997) considered the scheduling of a finite number of arrivals. Denton and Gupta (2003) conducted a numerical study using different

numbers of patients (e.g., 3, 5, and 7) to determine arrival intervals between patients and modeled service times using the uniform distribution. These models assume that a finite number of patients arrive at deterministic times and that their service involves an exponentially distributed duration. In contrast, we model surgery duration using the lognormal, which is regarded as a good fit; the gamma, which can be shaped similar to the lognormal; and the normal, which is used extensively because of its tractability and general applicability.

Weiss (1990) was the first to address the scheduling (i.e., time tabling) of two surgeries for a given sequence. His model prescribes the starting time of the second surgery with the objective of minimizing the sum of the expected costs of surgeon's waiting- and OR-idle- times. In contrast, we focus on sequencing of surgeries rather than scheduling starting times. Weiss (1990) showed that, if surgery times are *i.i.d.* and *symmetrical*, as is the normal distribution, for example, sequencing surgeries according to the SV rule is optimal. Gupta (2007) and Denton et al. (2007) used stochastic ordering to schedule two surgeries with durations that have the same mean but different variances. However, they cite no reference that indicates these relationships are prevalent in practice. Pinedo (2009) discussed the scheduling of two surgeries with durations that are independent and uniformly distributed, arguing that variance has a much stronger influence on the optimal schedule than does the mean. In contrast, we study applicable distributions rather than imposing restrictions such as symmetry or stochastic ordering.

Assuming that short surgery durations inherently exhibit less variability than long ones, Lebowitz (2003) studied the SM rule, using Monte Carlo simulation to show that it can improve on-time performance and decrease overtime expense. Sier et al. (1997) described a practice that sequences surgeries according to patient age and estimates of surgery durations, scheduling the younger patient first or using LM

if ages are the same. The rules proposed by Lebowitz (2003) and Sier et al. (1997) are based on experience or assumption, but our work provides analytical and numerical results.

Mathematical programming models formulated to prescribe surgery scheduling may be categorized as deterministic (Guinet and Chaabane, 2003; Jebali et al., 2006; Cardeon et al., 2009; Fei et al., 2008, 2009) or stochastic (Denton and Gupta, 2003; Denton et al., 2007; Lamiri et al., 2008, 2009). All approaches proposed to optimize the former all employ column generation to assign each surgery to a specific OR and to sequence surgeries in each OR each day; the latter employ some sampling method (e.g., the sampling average approximation) that uses a limited number of scenarios to represent the broad range of surgery-durations outcomes. Rather than developing a solution algorithm as in the deterministic case or attempting to represent a broad range of possible outcomes with a limited number of scenarios, we study stochastic sequencing policies both analytically and numerically, representing all possible outcomes.

3. CAPACITY ALLOCATION

This chapter proposes strategic-level models to allocate surgical specialties to operating room (OR) days, each defined as the capacity provided by one OR during a work day of duration h hours. The objective is to minimize total expected costs due to penalties for any patients who are not accommodated and for under- and over-usage of OR capacity, which result in under-utilization and overtime, respectively (Dexter et al., 2003). Such an allocation model provides a plan by which surgeons can schedule their activities, assuring that they can balance time performing surgery, time supporting office hours, and time fulfilling other responsibilities. The plan can also be used by hospital administrators to plan OR capacity, for example by initiating an expansion if an excessive number of patients can not be accommodated, and to integrate the efforts of surgery support staff (e.g., anesthesiologists and nurses) and ancillary departments such as a PACU and a SICU.

Strategic level decisions may involve capacity allocation (Denton et al., 2010; Zhang et al., 2009), capacity expansion (Lovejoy and Li, 2002), and patient-mix problems (Gupta and Wang, 2008; Zhang et al., 2009). We do not consider capacity expansion decisions, which are typically made over a time horizon of 3 - 5 years; rather, we focus on capacity allocation, which commonly deals with a one-year horizon and provides guidelines for subsequent tactical decisions that refine capacity allocations. This chapter regards patient-mix as a byproduct of capacity allocation decisions.

The remainder of this paper is organized as follows. Section 3.1 presents preliminaries and formulates our prototypical model. Section 3.2 describes our four model adaptations: NV-CA, NV-SIP, SIP, and NS-SIP. Section 3.3 presents a numerical ex-

periment that compares the computational efficacies of the four models, emphasizing run time and solution quality.

3.1 Preliminaries

This section presents preliminaries that underlie our study. Subsections discuss our assumptions, define the notation we use, and present our prototypical stochastic, non-linear optimization model (NL-CA), which allocates each surgical specialty to a specific number of OR days.

3.1.1 Assumptions

We assume that hospital administrators know the distribution functions for demand and surgery durations and that weekly demand is stationary over the planning horizon. For planning purposes, we form a representative duration (Choi and Wilhelm, 2012b) for each specialty and interpret it as the duration of a randomly selected surgery requiring this specialty. Each specialty performs hundreds of different procedures, which can be classified by current procedure terminology codes (CPT) and their combinations. We determine the representative duration for each specialty by forming the convex combination of the durations of these relevant procedures according to their (historical or forecast) frequencies (i.e., probabilities) of occurrence and then invoking the central limit theorem (Casella and Berger, 2001) to justify assuming that each is normally distributed.

We assume that at most one specialty is allocated to each OR each day, but we allow each specialty to be allocated to more than one OR and/or more than one day to accommodate demand. We assume that the number of accommodated patients requiring a particular specialty is the same on each OR day to which the specialty is allocated.

We partition ORs into subsets, each of which is outfitted with similar equipment

and can thus support the same types of surgeries. Similarly, we partition surgical specialties into subsets, each of which require the same OR equipment. This decomposes the overall problem into an index set $\mathcal{K}$ of independent and pairwise-disjoint problems, each $\kappa \in \mathcal{K}$ involving one subset of ORs and the subset of specialties that can be performed using the equipment they offer. The cost parameters associated with one subset of ORs may differ from those related to others because the equipment installed in them and the surgical specialists who use them may give rise to unique costs.

3.1.2 Notation

We use the following indices and sets in formulating our model:

Index Sets and Indices

M	ORs	$m \in M$
N	surgery specialties, (e.g., orthopedic, cardiovascular)	$n \in N$
$\mathcal{K}$	compatible surgical specialties and ORs	$\kappa \in \mathcal{K}$
M_κ	ORs dedicated to specialties $n \in N_\kappa$	
N_κ	specialties to be performed in ORs $m \in M_\kappa$	
D	days (e.g., Monday through Friday)	$d \in D = \{1, \ldots, 5\}$

Two types of random variables are associated with specialty $n \in N$: A_n denotes the forecast number of surgeries demanded each period (e.g., week); and P_n denotes the random, representative duration of each surgery. We assume that weekly demand A_n is Poisson distributed with mean rate λ_n. Let P_n be normally distributed with mean μ_n and variance σ_n^2. Let h denote the length of the standard OR workday (e.g., 8 hours).

Two types of decision variables allocate specialties to OR days and, as a by product, determine patient mix: R_n prescribes the number of OR days to which specialty $n \in N_\kappa$ is allocated, and V_n gives the number of representative surgeries

requiring specialty n that are assigned each day to each OR in set M_κ to which specialty n is allocated. Following earlier work Zhang et al. (2009), Testi et al. (2007), Adan and Vissers (2002), and Blake and Donald (2002), we use integer variables to prescribe allocation decisions at the strategic-level. We assume hospital administrators and surgeons can map the R_n solutions that our models prescribe to assign each specialty to particular day(s) of the week. Such assignments are likely to be heavily influenced by the schedule that has been used historically as well as the preferences of surgeons.

The random variable that describes the time to complete the set of surgeries allocated to an OR day is the sum of V_n *i.i.d.* P_n's, which we denote by $[V_n * P_n]$. We assign the same random workload, $[V_n * P_n]$, to each OR on each day to which specialty n is allocated.

Other random variables are related to cost penalties incurred by allocating V_n surgeries of specialty n to each OR day of duration h: $U_n = \max(h - [V_n * P_n], 0)$ defines the under-usage of each OR day, $O_n = \max([V_n * P_n] - h, 0)$ specifies the over-usage of each OR day, $\bar{A}_n = \max(A_n - R_n V_n, 0)$ gives the number of patients requiring specialty n who are not accommodated, and $\bar{S}_n = \min(A_n, R_n V_n)$ defines the number of patients requiring specialty n who are accommodated.

The objective of our model is to maximize *excess revenue*. For specialty n, we use Π_n to denote the excess revenue for each surgery performed; that is, the total reimbursement to the hospital minus its direct costs (e.g., operating staff, OR equipment, the OR facility, and overhead). From another perspective, Π_n is the excess revenue foregone if a surgery is not accommodated. If the surgeon has privileges at another hospital, s/he can take her/his patient elsewhere, so the surgeon would not be giving up his income and the patient would still receive surgery. In addition, we include a penalty of $\hat{c}_n^a$ for each surgery of specialty n that is not accommodated,

for example, representing the cost of administrative effort to handle the overload; the loss of good will, although the patient may not be perturbed if the surgery is performed at another hospital; and/or the loss of patient satisfaction due to delaying an elective surgery. If an emergency surgery could not be accommodated and could not be performed at another hospital, the penalty would be severe. The cost structure also depends upon the organizational structure: if the surgeon worked for the hospital and could not perform a surgery elsewhere, Π_n would have to be defined appropriately.

Other relevant cost parameters include penalties associated with specialty $n \in N_\kappa$ and OR subset $\kappa \in \mathcal{K}$: c_n^u for under-usage of OR time relative to h each day (i.e., idleness), and c_n^o for over-usage of OR time each day (i.e., overtime for any surgery time beyond h hours). The cost of underutilization, c_n^u, could account for the fixed cost of the equipment and facility itself. Alternatively, it could be considered the opportunity cost of foregoing excess revenue so that incorporating the factor $\left(\Pi_n/\mu_n\right)$ would cost underutilization (in terms of hours) more on a par with not accommodating a patient. Overtime is bad in that the surgery staff is paid a premium, which may depend upon union contract or hospital policy, increasing direct cost by an increment. However, it is good in that equipment and the OR are utilized when, otherwise, they would go idle. It is also good in that the hospital earns excess revenue for surgeries completed after time h, allowing more patients to be accommodated. Clearly, more surgeries could be accommodated on overtime, increasing excess revenue, at the cost of overtime premium, so these parameter values set the stage for interesting trade offs.

3.1.3 Prototypical Allocation Model (NL-CA)

Our prototypical formulation (NL-CA) represents issues relevant to allocating surgical specialties to OR-days in a succinct manner. The objective is to maximize excess revenue:

$$\max \quad \sum_{\kappa \in \mathcal{K}} \sum_{n \in N_\kappa} \Big\{ \Pi_n E[\bar{S}_n] - \hat{c}_n^a E[\bar{A}_n] - R_n \{ c_n^u E[U_n] + c_n^o E[O_n] \} \Big\}, \tag{3.1}$$

where, for specialty n, $E[\cdot]$ denotes the expected values of $\bar{S}_n$, $\bar{A}_n$, U_n, and O_n, respectively. Noting that $\bar{S}_n = \min(A_n, R_n V_n) = -\max(-A_n, -R_n V_n) = A_n - A_n - \max(-A_n, -R_n V_n) = A_n - \max(A_n - R_n V_n, 0) = A_n - \bar{A}_n$ we substitute $E[\bar{S}_n] = E[A_n] - E[\bar{A}_n]$, reforming the objective as

$$\max \quad \sum_{\kappa \in \mathcal{K}} \sum_{n \in N_\kappa} \Big\{ \Pi_n E[A_n] - c_n^a E[\bar{A}_n] - R_n \{ c_n^u E[U_n] + c_n^o E[O_n] \} \Big\}, \tag{3.2}$$

where $c_n^a = \Pi_n + \hat{c}_n^a$.

Because $\Pi_n E[a_n]$ is a constant, the prototypical allocation model can be formulated as

$$(NL-CA) \ \min \ \sum_{\kappa \in \mathcal{K}} \sum_{n \in N_\kappa} \Big\{ c_n^a E[\bar{A}_n] + R_n \{ c_n^u E[U_n] + c_n^o E[O_n] \} \Big\} \tag{3.3}$$

$$s.t. \sum_{n \in N_\kappa} R_n \leq |M_\kappa||D| \qquad \kappa \in \mathcal{K} \tag{3.4}$$

$$[V_n * P_n] + U_n - O_n = h \qquad \kappa \in \mathcal{K}, n \in N_\kappa \tag{3.5}$$

$$R_n V_n + \bar{A}_n \geq A_n \qquad \kappa \in \mathcal{K}, n \in N_\kappa \tag{3.6}$$

$$V_n \leq \alpha_1 \Big(\frac{h}{\mu_n}\Big) \qquad \kappa \in \mathcal{K}, n \in N_\kappa \tag{3.7}$$

$$R_n \leq \alpha_2 \lambda_n / \Big(\frac{h}{\mu_n}\Big) \qquad \kappa \in \mathcal{K}, n \in N_\kappa \tag{3.8}$$

$$V_n, R_n \in \mathcal{Z}^+ \qquad \kappa \in \mathcal{K}, n \in N_\kappa \tag{3.9}$$

$$\bar{A}_n, O_n, U_n \in \Re^+ \qquad \kappa \in \mathcal{K}, n \in N_\kappa. \tag{3.10}$$

Objective function (3.3) minimizes total expected costs due to penalties for any patients who are not accommodated (i.e., the first term), including revenue foregone as well as associated administrative costs, and for under- and over-usage of OR days (i.e., the last two terms), respectively. Constraints (3.4) ensure that total number of OR days allocated to all specialties $n \in N_\kappa$ cannot be larger than the total number of OR days available $|M_\kappa||D|$. Constraints (3.5) define the under- (U_n) and over- (O_n) usage of each OR day to which specialty n is allocated, given that $[V_n * P_n]$ denotes the workload associated with allocating V_n surgeries of specialty n to each OR and each day. Constraints (3.6) define the number of patients who are not accommodated ($\bar{A}_n$). Rather than using arbitrarily large upper bounds on integer decision variables V_n and R_n, (3.7) and (3.8) invoke practical bounds with the goal of managing model tightness to facilitate run time. The upper bound that constraint (3.7) imposes on decision variable V_n is a multiple (α_1) of the expected number of surgeries that can be accommodated in an OR day of duration h (i.e., h/μ). Similarly, the upper bound that constraint (3.8) imposes on decision variable R_n is a multiple (α_2) of the expected number of OR days required by surgical specialty n, determined

as the expected demand divided by the expected number of patients that can be accommodated each OR day. Specifying values of $\alpha_1 = \alpha_2 = 2$ would be reasonable. Constraints (3.9) and (3.10) impose integer requirements and sign restrictions on decision variables. To facilitate presentation, we do not repeat restrictions (3.7)-(3.10) in following models.

Clearly, NL-CA is separable on κ, leading to $|\mathcal{K}|$ independent problems, NL-CA$_\kappa$. The primary advantages of this model are that it presents the fundamental issues involved in a succinct manner and that it demonstrates separability relative to κ. Subsequent models (i.e., NV-CA, NV-SIP, NV-SIP and NS-SIP) relate to such decomposed sub-problems. The primary disadvantage of model NL-CA is that it involves nonlinearities; products of decision variables in the objective function and in constraints (3.6); and the form $[V_n * P_n]$, which represents the sum of V_n *i.i.d.* random variables.

In the context of a two-stage stochastic program, R_n and V_n are prescribed in the first stage before demands and durations are realized. Once these uncertainties have been resolved, decision variables $\bar{A}_n$, U_n, and O_n can be prescribed.

3.2 Model Adaptations and Solution Methods

In this section, we propose four adaptations of prototypical model NL-CA. The goal of this study is to overcome the disadvantages of the prototype by deriving a linear form that facilitates solution. Each of these models offers advantages but is also attended by disadvantages. Each adaptation may be better suited to particular applications. Also, each adaptation may be better suited to a different solution approach.

Each of the following four subsections formulates one of the model adaptations; describes relevant advantages, disadvantages, and applications; and outlines a solu-

tion approach. The first two models, NV-CA and NV-SIP, exploit the relationship of constraints (3.5) to the news vendor problem, dealing with the nonlinear $[V_n * P_n]$ term to obtain the expected values of underage, $E[U^*_\kappa]$, and overage, $E[O^*_\kappa]$. The third model, SIP, recasts general integer decision variables V_n and R_n to allocate specialty n to specific ORs $m \in M_\kappa$ and days $d \in D$, adding subscripts m and d, to obtain V_{nmd} and R_{nmd} and defining the later as a binary decision variable. It also replaces the nonlinear term $[V_n * P_n]$ with the product of decision variable V_{nmd} and random variable P_n. The fourth model, NS-SIP, incorporates linear forms to eliminate nonlinear terms, each formed by the product of two decision variables, and symmetry induced in model SIP relative to m and d.

The solution approach that we propose for each of the last three models involves a two-stage stochastic integer program, giving rise to the SIP designation in each acronym. We employ stochastic programming by generating a set of scenarios to evaluate $E[O_n], E[U_n]$ and $E[\bar{A}_n]$. A scenario is one specific, complete realization of the stochastic elements that underlie the formulation (e.g., actual demand and representative duration for each specialty). We model uncertain surgery demand and duration via the multi-variate random variable $\tilde{\omega}$ for which a realization is commonly referred to as a scenario. For each scenario $\omega \in \Omega$, we define a problem, which we refer to as the recourse problem. We discretize probability measure function $q^\omega = \mathscr{P}(\tilde{\omega} = \omega)$ for each scenario $\omega \in \Omega$ and evaluate the deterministic-equivalent form of the stochastic program (Birge and Louveaux, 1997) to prescribe solutions to models NV-SIP, SIP, and NS-SIP.

In large-scale problems, the number of scenarios is huge and the possibility of using statistical estimations of the recourse function becomes computationally attractive. The simplest method for incorporating statistical approximations in solution procedures is to replace the recourse function, $g(V, R, \omega)$, with a sample mean

approximation. We solve the sample mean problem with a collection of independent and identically distributed observations of $\tilde{\omega}$.

3.2.1 News Vendor-based Capacity Allocation (NV-CA)

Prototype model NL-CA aggregates sub-specialties within each specialty to form the representative duration of each surgery associated with the specialty. Model NV-CA further aggregates specialties $n \in N_\kappa$, forming representative duration P_κ for all specialties $n \in N_\kappa$ in a manner analogous to the one we used to form P_n for specialty n. Since representative durations are used, $\bar{A}, U, O$, and V are associated with set κ, not specialty n, and become $\bar{A}_\kappa, U_\kappa, O_\kappa$, and V_κ. Representative surgeries are then allocated to each day to each OR in set M_κ. Decomposed subproblem κ provides a capacity of $|M_\kappa| \times |D|$ OR days. Our model for decomposed problem κ, (NV-CA$_\kappa$), is based on the premise that V_κ surgeries of type κ are allocated to each OR $m \in M_\kappa$ each day $d \in D$ and is

$$(NV - CA_\kappa) \qquad \min \quad c^a_\kappa E[\bar{A}_\kappa] + |M_\kappa||D|\{c^u_\kappa E[U_\kappa] + c^o_\kappa E[O_\kappa]\} \tag{3.11}$$

$$s.t. \quad [V_\kappa * P_\kappa] + U_\kappa - O_\kappa = h \tag{3.12}$$

$$|M_\kappa||D|V_\kappa + \bar{A}_\kappa \geq A_\kappa. \tag{3.13}$$

Constraints like (3.4) in NL-CA are not needed because NV-CA eliminates R_κ decision variables; only one aggregated specialty is assigned to subset κ and uses its capacity $|M_\kappa||D|$ exclusively. Constraints (3.12) ((3.13)) are equivalent to (3.5) ((3.6)).

The primary advantage of NV-CA is that it simplifies NL-CA, eliminating non-linear terms, each formed by the product of two decision variables. This advantage of simplicity also predisposes its applicability to cases in which it is meaningful to

assign aggregated specialties to each OR in subset $\kappa \in \mathcal{K}$ each day; we advocate it for what we call rough-cut capacity planning, a process that assesses whether a hospital has enough overall OR capacity to deal with future demand or not.

NV-CA retains the $[V_n * P_n]$ term, but it is amenable to a straightforward heuristic based on the *inverse news vendor* model. Consider a subproblem based on constraints (3.12) with the objective of minimizing the sum of expected under- and over-usage; expressed in the form of a news vendor problem, we have

$$\min_{V_\kappa} \quad c_\kappa^u E[(h - [V_\kappa * P_\kappa])^+] + c_\kappa^o E[([V_\kappa * P_\kappa] - h)^+]. \tag{3.14}$$

Assuming that the duration of the OR day is fixed to be h and that aggregated durations durations $P_\kappa, \kappa \in \mathcal{K}$, are normally distributed (Choi and Ketzenberg, 2012), the optimal number of surgeries to allocate, V_κ^*, can be found from

$$F_{[V_\kappa * P_\kappa]}(h) = \frac{c_\kappa^o}{c_\kappa^o + c_\kappa^u} = \Phi(z), \tag{3.15}$$

where $F_{[V_\kappa * P_\kappa]}$ is the distribution function of $[V_\kappa * P_\kappa]$. V_κ^* can be expressed explicitly as follows:

$$\begin{aligned} V_\kappa^* &= \llcorner \hat{V}_\kappa \lrcorner \text{ or } \ulcorner \hat{V}_\kappa \urcorner, \\ \text{where } \hat{V}_\kappa &= \Big(\frac{-z\sigma_\kappa + \sqrt{z^2\sigma_\kappa^2 + 4\mu_\kappa h}}{2\mu_n} \Big)^2. \end{aligned}$$

We can incorporate V_κ^* into constraint (3.13) and use it to determine the associated value of decision variable $\bar{A}_\kappa$. Defining $E[U_\kappa^*] := E[(h - [V_\kappa^* * P_\kappa])^+]$ and $E[O_\kappa^*] := E[([V_\kappa^* * P_\kappa] - h)^+]$ (see (Choi and Wilhelm, 2012b)), the values of V_κ^*, $\bar{A}_\kappa$, $E[U_\kappa^*]$, and $E[O_\kappa^*]$ can be incorporated in objective function (3.11) to obtain the

solution value that this heuristic prescribes.

V_κ^* is optimal with respect to the news vendor problem but may not lead to globally best possible values for $\bar{A}_\kappa$, $E[U_\kappa^*]$, and $E[O_\kappa^*]$. For that reason, this method must include a local search on V_κ values (e.g., .., $V_\kappa^* - 1$, $V_\kappa^* + 1$, ...).

3.2.2 News Vendor-based Heuristic (NV-SIP)

Model NV-SIP retains the focus of the prototype model, NL-CA, seeking to prescribe V_n^*, the optimal number of patients requiring specialty $n \in N_\kappa$ each OR day, but adopting the *inverse news vendor* model proposed in the previous section. Subsequently, we solve a stochastic integer programming with V_n^* fixed to determine the number of OR days for each specialty, R_n^*.

Consider a subproblem based on constraints (3.12) with the objective of minimizing the sum of expected under- and over-usage; expressed in the form of a news vendor problem, we have

$$\min_{V_n} \quad c_n^u E[(h - [V_n * P_n])^+] + c_n^o E[([V_n * P_n] - h)^+]. \tag{3.16}$$

Let V_n^* be the optimal solution to (3.16), and define $E[U_n^*] := E[(h - [V_n^* * P_n])^+]$ and $E[O_n^*] := E[([V_n^* * P_n] - h)^+]$. Next, we adapt prototype model NL-CA, relaxing constraints (3.5), (3.7) and (3.8) and incorporating values V_n^*, $E[U_n^*]$ and $E[O_n^*]$. The NV-SIP model for subset κ is a two-stage stochastic program that prescribes decision variables R_n^* and $E[\bar{A}_\kappa^*]$.

$$(NV - SIP_\kappa) \qquad \min \quad \sum_{n \in N_\kappa} R_n \{c_n^u E[U_n^*] + c_n^o E[O_n^*]\} + \sum_{\omega \in \Omega} q^\omega h_\kappa(R, \omega) \tag{3.17}$$

$$s.t. \quad (3.4).$$

For each scenario ω,

$$\min \quad h_\kappa(R,\omega) = \sum_{n \in N_\kappa} c_n^a \bar{A}_n^\omega \tag{3.18}$$

$$s.t. \quad \bar{A}_n^\omega \geq A_n^\omega - R_n V_n^* \qquad\qquad n \in N_\kappa. \tag{3.19}$$

NV-SIP offers a number of advantages. It simplifies NL-CA, eliminating nonlinear terms, each formed by the product of two decision variables, and deals with the $[V_n * P_n]$ term, which is amenable to our heuristic based on the *inverse news vendor* model. With V_n^* determined as the solution to an inverse news vendor problem, NV-SIP involves only the small number of $2|N|$ integer decision variables, so it could be used to solve the problem over all $\kappa \in \mathcal{K}$ without decomposing relative to κ. Finally, because NV-SIP retains the focus on individual surgery specialties instead of aggregated specialties as in NV-CA, it can be expected to find wider application. Like the NV-CA model, NV-SIP requires a search on $V_\kappa*$ values to guarantee the best possible global solution.

3.2.3 Stochastic Integer Programming (SIP)

Model SIP linearizes the prototype model. It adapts NL-CA by replacing the sum of random variables $[V_n * P_n]$ with the product of decision variable V_n and random variable P_n. Both forms have the same mean, $V_n * \mu_n$, but the variance of the former is $V_n * \sigma_n^2$ and that of the latter is $V_n^2 * \sigma_n^2$. Thus, this replacement introduces an additional variability due to multiplying by a constant.

Model SIP also replaces decision variables V_n and R_n by V_{nmd} and R_{nmd}, respectively, adding subscripts m and d to specialize each allocation of specialty n to a particular OR m and day d. At the same time, we revise general integer variable R_n to the binary form R_{nmd}, which is 1 if specialty n is allocated to OR m on day d, 0

otherwise. R_n in NL-CA is equivalent to $\sum_{m,d} R_{nmd}$ in SIP. Similarly, we revise U_n and O_n to represent under- and over-usage of OR m on day d, changing subscripts from n to m and d, to obtain U_{md} and O_{md}, respectively. With these adaptations, we propose the following two-stage stochastic integer linear formulation, SIP for subset κ:

$$(SIP_\kappa) \qquad \min \quad \sum_{\omega \in \Omega} q^\omega g_\kappa(V, R, \omega) \tag{3.20}$$

$$s.t. \quad \sum_{n \in N_\kappa} R_{nmd} \leq 1 \qquad m \in M_\kappa, d \in D \tag{3.21}$$

$$R_{nmd} \leq V_{nmd} \qquad n \in N_\kappa, m \in M_\kappa, d \in D \tag{3.22}$$

$$R_{nmd} \alpha \frac{h}{\mu_n} \geq V_{nmd} \qquad n \in N_\kappa, m \in M_\kappa, d \in D. \tag{3.23}$$

For each scenario ω,

$$\min \quad g_\kappa(V, R, \omega) = \sum_{n \in N_\kappa} c_n^a \bar{A}_n + \sum_{m \in M_\kappa} \sum_{d \in D} \{c_\kappa^u U_{md}^\omega + c_\kappa^o O_{md}^\omega\} \tag{3.24}$$

$$s.t. \quad \sum_{n \in N_\kappa} V_{nmd} p_n^\omega + U_{md}^\omega - O_{md}^\omega = h \qquad m \in M_\kappa, d \in D \tag{3.25}$$

$$\sum_{m \in M_\kappa} \sum_{d \in D} V_{nmd} + \bar{A}_n^\omega \geq A_n^\omega \qquad n \in N_\kappa. \tag{3.26}$$

Constraints (3.21) ensure that at most one specialty is assigned to each OR each day; (3.21) is equivalent to (3.4) in NL-CA. Constraints (3.22) and (3.23) ensure that the number of assigned patients is positive if and only if specialty n is assigned to OR m on day d, else both R_{nmd} and V_{nmd} must be zero. Owing to (3.22) and (3.23),

constraints (3.5) and (3.6) of NL-CA can be recast in the linearized forms (3.25) and (3.26). Constraints (3.25) linearize constraints (3.5) of NL-CA, replacing $[V_n * P_n]$ with the product $V_{nmd}P_n^{\omega}$ for each scenario ω.

The primary advantages of SIP is that it linearizes NL-CA with the expectation that it will facilitate run time and that it focuses on individual specialties n rather than on aggregations of specialties as in NV-CA. Its primary disadvantages are that it introduces additional variability due to replacing $[V_n * P_n]$ with the product $V_{nmd}P_n^{\omega}$ and that it introduces a high degree of symmetry relative to m and d in replacing decision variables V_n and R_n with V_{nmd} and R_{nmd}, respectively, to effect the linearization of product terms $R_n * V_n$ in (3.6).

3.2.4 Stochastic Integer Programming without Symmetry

Model NS-SIP seeks to improve model SIP by eliminating the symmetry relative to m and d (SIP entails $\sum_{\kappa} |N_{\kappa}||M_{\kappa}||D|$ general integer variables V_{nmd} and binary variables R_{nmd}) with the goal of facilitating run time. To avoid symmetry, we introduce new decision variables and constraints. The new index set $S_{\kappa} = \{1, \ldots, |M_{\kappa}||D|\}$ denotes the number of OR days to which specialties $n \in N_{\kappa}$ can be allocated, where $|M_{\kappa}||D|$ is the total number of OR days (i.e., capacity) available. New binary decision variable R_{ns}^B is 1 if specialty n is allocated to $s \in S_{\kappa}$ OR days, 0 otherwise.

To linearize product forms V_nR_n, $R_nU_n^{\omega}$, and $R_nO_n^{\omega}$, we introduce RV_n, RU_n^{ω}, and RO_n^{ω}. We define sufficiently large numbers $\bar{V}_n, \bar{U}_n$, and $\bar{O}_n$ that exceed upper limits on RV_n, RU_n^{ω}, and RO_n^{ω}, respectively. For example, $\bar{V}_n = |M_{\kappa}||D| * h/\mu_n$. Following this introduction, we present model (NS-SIP), a stochastic integer linear program

that avoids symmetry:

$$(NS-SIP_\kappa) \qquad \min \quad \sum_{\omega\in\Omega} q^\omega g_\kappa^{NS}(V,R,R^B,RV,\omega) \tag{3.27}$$

$$s.t. \qquad (3.4)$$

$$\sum_{s\in S_\kappa} R^B_{ns} = 1 \qquad n\in N_\kappa \tag{3.28}$$

$$R_n = \sum_{s\in S_\kappa} sR^B_{ns} \qquad n\in N_\kappa \tag{3.29}$$

$$RV_n \geq sV_n - s(1-R^B_{ns})\bar{V}_n \qquad n\in N_\kappa, s\in S_\kappa. \tag{3.30}$$

$$RV_n \leq sV_n + s(1-R^B_{ns})\bar{V}_n \qquad n\in N_\kappa, s\in S_\kappa. \tag{3.31}$$

For each scenario ω,

$$\min \quad g_\kappa^{NS}(V,R,R^B,RV,\omega) = \sum_{n\in N_\kappa}\{c^a_n\bar{A}_n + c^u_n RU^\omega_n + c^o_n RO^\omega_n\} \tag{3.32}$$

$$s.t. \quad V_n p^\omega_n + U^\omega_n - O^\omega_n = h \qquad s\in S_\kappa \tag{3.33}$$

$$RV_n + \bar{A}^\omega_n \geq a^\omega_n \qquad n\in N_\kappa \tag{3.34}$$

$$RU^\omega_n \geq sU_n - s(1-R^B_{ns})\bar{U}_n \qquad n\in N_\kappa, s\in S_\kappa \tag{3.35}$$

$$RU^\omega_n \leq sU_n + s(1-R^B_{ns})\bar{U}_n \qquad n\in N_\kappa, s\in S_\kappa \tag{3.36}$$

$$RO^\omega_n \geq sO_n - s(1-R^B_{ns})\bar{O}_n \qquad n\in N_\kappa, s\in S_\kappa \tag{3.37}$$

$$RO^\omega_n \leq sO_n + s(1-R^B_{ns})\bar{O}_n \qquad n\in N_\kappa, s\in S_\kappa. \tag{3.38}$$

Constraints (3.28) ensure that specialty n is allocated to a specific number, s, of

OR days. Constraints (3.29) define R_n using binary decision variables R^B_{ns} (i.e., $R_n = s$ if $R^B_{ns} = 1$). Constraints (3.30) and (3.31) ensure that $RV_n = sV_n = R_nV_n$ if $R^B_{ns} = 1$. If $R^B_{ns} = 0$, constraints (3.30) and (3.31) are redundant. Similarly, constraints (3.35) and (3.36) ensure that $RU^\omega_n = sU^\omega_n = R_nU^\omega_n$ if $R^B_{ns} = 1$. If $R^B_{ns} = 0$, constraints (3.35) and (3.36) are redundant. Invoking the same logic, constraints (3.37) and (3.38) ensure that $RO^\omega_n = sO^\omega_n = R_nO^\omega_n$ if $R^B_{ns} = 1$. If $R^B_{ns} = 0$, constraints (3.37) and (3.38) are redundant. Actually, the effect of the minimizing objective function renders (3.36) and (3.38) unnecessary and we do not include them in the test cases reported in the next section. Both constraints (3.30) and (3.31) are, however, necessary.

The primary advantages of NS-SIP are that it is a linear model that avoids symmetry with the goal of facilitating solution and that it focuses on individual specialties n to promote applicability. The main disadvantage of NS-SIP is that it requires large numbers of constraints to effect linearization and avoidance of symmetry (i.e., (3.35)-(3.38)).

3.3 Computational Evaluation

The goals of our experiment are to compare our four models in terms of problem size, solution quality, and run time and to evaluate the performance of the sample mean approach in application to our three SIP models by generating small, medium and large numbers of scenarios - 50, 150 and 250. This section describes our experiment in two subsections. The first describes the design of our experiment and the second reports test results.

3.3.1 Design of Experiment

To evaluate our three SIP models, we employ four factors, three each with three levels and one with two levels, creating a total of 54 cases. We denote the first factor,

which determines problem size, using ($|N|$, $|M|$) to indicate that any of $|N|$ specialties can use any of $|M|$ ORs. The three levels we use are (5, 5), (5,5)+(5,5), and (10, 10), where the second can be decomposed into two independent (5, 5) problems and the third is a relaxation of the second that allows each of 10 specialties to use any of the 10 ORs. These levels (small, medium, large) provide a good test bed to evaluate our models and represent the sizes of many actual problems. The second factor is the coefficient of variation (CV) of surgery duration for which we use two levels: 0.1 and 0.7. The third factor is the model and our three levels are (NV-SIP, SIP, NS-SIP). The fourth factor is the number of scenarios for which we use three levels: (50, 150, 250). For each case, composed by selecting one level of each factor, we solve 20 independent replications and table average performance measures.

In contrast, we solve NV-CA analytically for each ($|N|$, $|M|$) and CV combination. The number of integer decision variables for NV-CA is equal to the number of decomposed sets, $|\mathcal{K}|$.

Table 3.1 displays the size of each problem. The seven columns give, respectively, ($|N|$, $|M|$), model, number of scenarios, numbers of variables (continuous, general integer (GI), binary integer (BI)) and constraints. Model NV-SIP has the smallest numbers of variables and constraints because we determine values of GIs, V_n, analytically. Model NS-SIP has fewer GIs than SIP but more constraints.

We generate λ_n, the mean rate of demand for surgeries associated with specialty n, employing $U[10, 50]$ and each random representative surgery duration employing a normal distribution with mean value from $U[0.5, 4.5]$ and CV from $\{0.1, 0.7\}$. We chose these ranges of parameter values by reviewing other papers to assure that they represent a range of actual cases (Beliën and Demuelemeester, 2006; Fei et al., 2010; Zhang et al., 2009). To provide a relative basis for comparison, we fix $c_u = 1$ and generate cost-parameter ratio c^o/c^u from $U[0.5, 1.5]$. Because c_n^a is a penalty per

Table 3.1: Problem Size

($\|N\|$, $\|M\|$)	Model	Scenarios	Continuous	GI	BI	Constraints
(5,5)	NV-SIP	50	250	5	-	251
		150	750	5	-	751
		250	1,250	5	-	1,251
	SIP	50	2,750	125	125	1,775
		150	8,250	125	125	4,775
		250	13,750	125	125	7,775
	NS-SIP	50	1,050	5	125	12,890
		150	3,150	5	125	37,890
		250	5,250	5	125	62,890
(5,5)×2	NV-SIP	50	500	10	-	502
		150	1,500	10	-	1,502
		250	2,500	10	-	2,502
	SIP	50	2,250	250	250	3,550
		150	8,250	250	250	9,550
		250	13,750	250	250	15,550
	NS-SIP	50	2,050	10	250	25,780
		150	6,150	10	250	75,780
		250	10,250	10	250	125,780
(10,10)	NV-SIP	50	500	10	-	501
		150	1,500	10	-	1,501
		250	2,500	10	-	2,501
	SIP	50	5,500	500	500	4,050
		150	16,500	500	500	10,050
		250	27,500	500	500	16,000
	NS-SIP	50	2,050	10	500	51,530
		150	6,150	10	500	151,530
		250	10,250	10	500	251,530

surgery and c^u is the cost per hour, we generate c^a using $c_n^a = c_u * \mu_n * U[1.7, 3.3]$.

In all tests, we run CPLEX 12.1 with default settings on a 3.00 GHz CPU of Intel Core Quad with 8 GB RAM. We use a time limit of 3,600 seconds (i.e., 1 hour) for each run, noting that preliminary tests have shown that using a limit of 2, or even 3, hours does not allow for significantly better convergence.

3.3.2 Test Results

Table 3.2 details results for all cases. The eight columns give, respectively, problem size ($|N|$, $|M|$), CV, model, number of scenarios, objective function mean and variance (over 20 replications), GAP, and CPU run time (seconds). We define GAP = 100 (ZI - ZL)/ZI, where ZI = current incumbent IP solution value and ZL= current relaxed LP solution value. Note that ZL can increase over time as CPLEX adds

cuts; at the optimal solution, GAP = 0 because ZI = ZL. The objective function values achieved by SIP and NS-SIP are comparable, but that of NV-SIP is smaller. Objective function values depend upon the variance of surgery duration (Choi and Wilhelm, 2012b) and SIP and NS-SIP have larger variability than NV-SIP, as discussed in section 3.2.3. Hence, NV-SIP gives a narrower confidence interval for the objective function value.

As the number of scenarios increases, mean values of objective function estimates do not change appreciably but variances tend to be smaller. However, the run time required to converge increases with the number of scenarios.

In general, run time tends to increase with CV. For example, for problem size ($|N|$, $|M|$)= (5, 5) with CV = 0.1, SIP can converge within 100 seconds, but when CV is increased to 0.7, SIP cannot converge within 1 hour. Run time is substantially affected by the model used.

In the attempt to obtain convergence, we extended our run time limit from 1 hour to 10 hours. This longer run time achieved better GAP values but not convergence.

To foster further insights, we now compare solutions prescribed by the four models, basing discussion on Table 3.3, which gives results for a typical case with ($|N|$, $|M|$)=(5,5); CV = 0.1; and, for each of the three SIP models, 250 scenarios. We also focus on the rate of convergence using Figures 3.1 and 3.2. Run times for the NV-CA model, which we solve numerically using Excel, are negligible and are, therefore, not tabled; those for the other three models are given in Table 3.2.

Rows in Table 3.3 give prescribed values of the objective function value, $c_n^a E(\bar{A}_n)$, $c_n^a E(S_n)$, $c_n^u E(U_n)$, $c_n^o E(O_n)$, λ_n, $E[\bar{A}_n]$, $E(S_n)$, $E[U_n]$, and $E[O_n]$, V_n and R_n, respectively. We apply c_n^a to $E[S_n]$ in our analysis, because we do not generate individual values for Π_n or $\hat{c}_n^a$. Rows use "/" separator to report values prescribed for each of the five specialties (i.e., $n = 1, ..., 5$). Columns in Table 3.3 display the solution pre-

Table 3.2: Detail Comparison of NV-SIP, SIP, and NV-SIP Approaches

($\|N\|$, $\|M\|$)	CV	Model	Scenarios	Obj(mean)	Obj(stdev)	GAP	Run time
(5,5)	0.1	NV-SIP	50	53.71	2.09	0.0	0.02
			150	54.32	1.3	0.0	0.04
			250	54.22	0.8	0.0	0.07
		SIP	50	58.50	2.2	0.0	5.53
			150	59.14	1.56	0.0	27.59
			250	59.44	0.93	0.0	89.9
		NS-SIP	50	59.80	2.19	0.0	23.2
			150	59.47	1.54	0.0	149.4
			250	59.54	0.97	0.0	567.3
	0.7	NV-SIP	50	98.68	2.18	0.0	0.02
			150	98.86	0.99	0.0	0.03
			250	98.03	0.93	0.0	0.05
		SIP	50	134.30	5.45	3.85	3600
			150	135.10	2.97	4.71	3600
			250	135.80	1.28	5.32	3600
		NS-SIP	50	136.10	6.70	0.0	157.9
			150	136.09	3.13	0.0	1770.3
			250	135.49	2.06	7.12	3600
(5,5)×2	0.1	NV-SIP	50	140.91	2.84	0.0	0.02
			150	141.71	1.67	0.0	0.07
			250	142.74	1.01	0.0	0.16
		SIP	50	152.45	3.06	0.03	3600
			150	152.25	2.02	0.06	3600
			250	152.34	1.38	0.08	3600
		NS-SIP	50	152.40	4.22	5.21	3600
			150	152.90	1.72	7.62	3600
			250	153.80	1.66	8.19	3600
	0.7	NV-SIP	50	247.76	3.09	0.0	0.02
			150	245.78	1.86	0.0	0.07
			250	246.47	1.74	0.0	0.15
		SIP	50	316.55	7.19	7.59	3600
			150	314.81	3.95	8.45	3600
			250	316.11	2.69	8.82	3600
		NS-SIP	50	317.04	7.65	8.63	3600
			150	316.01	4.12	9.51	3600
			250	318.82	4.08	13.42	3600
(10,10)	0.1	NV-SIP	50	141.54	2.86	0.0	0.02
			150	142.10	2.07	0.0	0.08
			250	141.67	1.27	0.0	0.18
		SIP	50	151.53	4.27	0.31	3600
			150	152.12	1.72	0.59	3600
			250	151.88	1.57	1.03	3600
		NS-SIP	50	151.75	3.96	6.32	3600
			150	152.48	1.74	8.93	3600
			250	153.85	1.52	9.56	3600
	0.7	NV-SIP	50	244.27	2.55	0.0	0.02
			150	243.34	2.07	0.0	0.09
			250	244.62	1.61	0.0	0.12
		SIP	50	310.21	8.93	4.98	3600
			150	313.95	4.33	5.39	3600
			250	314.79	2.42	5.43	3600
		NS-SIP	50	315.84	7.56	9.12	3600
			150	317.98	5.21	11.76	3600
			250	319.27	4.47	15.12	3600

scribed by each of the four models (NV-CA, NV-SIP, SIP, and NS-SIP. The NV-CA column gives values associated with the aggregated model with $\kappa \in \mathcal{K}$ where $|\mathcal{K}| = 1$; each of the other columns gives results relative to specialty $n \in N = \{1, ..., 5\}$. Note that, because $c_n^u = 1$, $c_n^u E[U_n] = E[U_n]$. Also, because c_n^o is close to 1 and numbers are rounded, the values of $c_n^o E[O_n]$ and $E[O_n]$ are similar.

The models tend to fall in two groups with respect to the objective function values they prescribe: NV-CA and NV-SIP give similar values that are lower than the similar values reported by SIP and NS-SIP. We conjecture that this difference is caused by the variability that each of the models ascribes to surgery duration. In forming the convex combination the representative durations of relevant specialties, NV-CA multiplies each by a fraction less than one (i.e., $\lambda_n / \sum_{n \in N} \lambda_n$), and squaring that fraction to determine variance tends to reduce the variability ascribed to the aggregate, representative duration. Correspondingly, the objective function value reported by NV-CA is somewhat less than that of NV-SIP. In replacing the sum of random variables $[V_n * P_n]$ with the product of decision variable V_n and random variable P_n, SIP and NS-SIP increase the variability ascribed to representative surgery duration. Correspondingly, they report larger objective function values.

Table 3.3: Comparison of NV-SIP, SIP, and NV-SIP for ($|N|$, $|M|$)=(5,5) with CV=0.1 and 250 Scenarios

Model	NV-CA	NV-SIP	SIP	NS-SIP
Obj Value	53.63	55.02	60.79	59.61
$c_n^a E[\bar{A}_n]$	46.79	17.24/6.36/3.98/12.60/1.16	17.32/6.95/3.52/11.65/1.34	16.42/6.58/3.78/11.95/1.26
$c_n^a E[S_n]$	206.41	33.46/44.24/45.42/53.90/34.84	33.38/43.65/45.88/54.85/34.66	34.28/44.02/45.62/54.55/34.74
$c_n^u E[U_n]$	0.28	0.49/0.05/0.02/0.72/0.28	0.62/0.20/0.16/0.88/0.45	0.59/0.18/0.16/0.86/0.49
$c_n^o E[O_n]$	0.02	0.07/0.44/0.42/0.01/0.01	0.12/0.58/0.63/0.04/0.22	0.14/0.51/0.61/0.06/0.18
λ_n	154	13/23/38/35/45		
$E[\bar{A}_n]$	28.46	4.42 /2.89/3.06/6.63/1.45	4.44/3.16/2.71/6.13/1.67	4.21/2.99/2.91/6.29/1.58
$E[S_n]$	125.54	8.58/20.11/34.94/28.37/43.55	8.56/19.84/35.29/28.87/43.33	8.79/20.01/35.09/28.71/43.42
$E[U_n]$	0.28	0.49/0.05/0.02/0.72/0.28	0.62/0.20/0.16/0.88/0.45	0.59/0.18/0.16/0.86/0.49
$E[O_n]$	0.02	0.07/0.54/0.38/0.01/0.01	0.13/0.71/0.57/0.04/0.20	0.15/0.62/0.55/0.05/0.16
V_n	5	2/4/7/4/11	2/4/7/4/11	2/4/7/4/11
R_n	25	4/5/5/7/4	4/5/5/7/4	4/5/5/7/4

$E[\bar{A}_n]$, $E[U_n]$, or $E[O_n]$ values also depend upon variability and, thus, the model. In application to the data generated in our tests, the three SIP models prescribe three categories of $E[U_n]$, $E[O_n]$, and $E[\bar{A}_n]$, reflecting the tradeoffs involved:

(1) large (≥ 0.5) $E[U_n]$ and small (≤ 0.15) $E[O_n]$, $E[\bar{A}_n]$ is large (≥ 4.0) - see n =1, 4;

(2) small (≤ 0.2) $E[U_n]$ and large (≥ 0.37) $E[O_n]$, $E[\bar{A}_n]$ is small (≤ 3.2) - see n =2, 3;

(3) medium $E[U_n]$ and small (≤ 0.20) $E[O_n]$, $E[\bar{A}_n]$ is medium (1.5) - see n =5.

As $E[O_n]$ gets smaller and $E[U_n]$ gets larger, $E[\bar{A}_n]$ gets larger; similarly, as $E[O_n]$ gets larger and $E[U_n]$ gets smaller, $E[\bar{A}_n]$ gets smaller, as can be expected.

NV-CA reports values that represent weighted (i.e., $\lambda_n / \sum_{n \in N_\kappa} \lambda_n$) average values for the five specialty values reported by NV-SIP. Still, we see that the models give consistent results:

$$
\begin{aligned}
c^a_\kappa E[\bar{A}_\kappa] &\approx \sum_{n \in N_\kappa} c^a_n E[\bar{A}_n] \\
c^a_\kappa E[S_\kappa] &\approx \sum_{n \in N_\kappa} c^a_n E[S_n] \\
c^u_\kappa E[U_\kappa] &\approx \sum_{n \in N_\kappa} \frac{\lambda_n}{\sum_{n \in N_\kappa} \lambda_n} c^u_n E[U_n] \\
c^o_\kappa E[O_\kappa] &\approx \sum_{n \in N_\kappa} \frac{\lambda_n}{\sum_{n \in N_\kappa} \lambda_n} c^o_n E[O_n] \\
V_\kappa &\approx \frac{\lambda_n}{\sum_{n \in N_\kappa} \lambda_n} V_n \\
R_\kappa &= |M_\kappa||D| = \sum_{n \in N_\kappa} R_n.
\end{aligned}
$$

Table 3.3 shows that all three SIP models (NV-SIP, SIP, and NS-SIP) prescribe the same V_n and R_n values. NV-CA prescribes a single value for V_κ that is the convex combination of the five V_n values prescribed by each of the three SIP models

as indicated above. Similarly, NV-CA reports a single value for R_κ that is the sum of the R_n values prescribed by each of the three SIP models as shown above. Thus, all four models are consistent in prescribing values for these core allocation decision variables.

Figure 3.1 (3.2) demonstrates how the value of the objective function changes over time as measured by GAP. Both figures show results for the case with ($|N|$, $|M|$)=(10,10), CV = 0.1, 250 scenarios, and models SIP and NS-SIP. We selected problem size ($|N|$, $|M|$)=(10,10) for this analysis because it defines the largest cases and takes longest to converge. NS-SIP makes a rapid improvement in the objective function value initially but SIP gives comparable values not long afterwards; both converge slowly after about 500 seconds. The GAP achieved by SIP decreases rapidly to less than 1% after about 1,000 seconds, but the GAP attained by NS-SIP remains large, even though these two models give similar objective function values. This result occurs because the NS-SIP model is not tight; the value of its linear relaxation is zero initially and increases very slowly as CPLEX adds cuts.

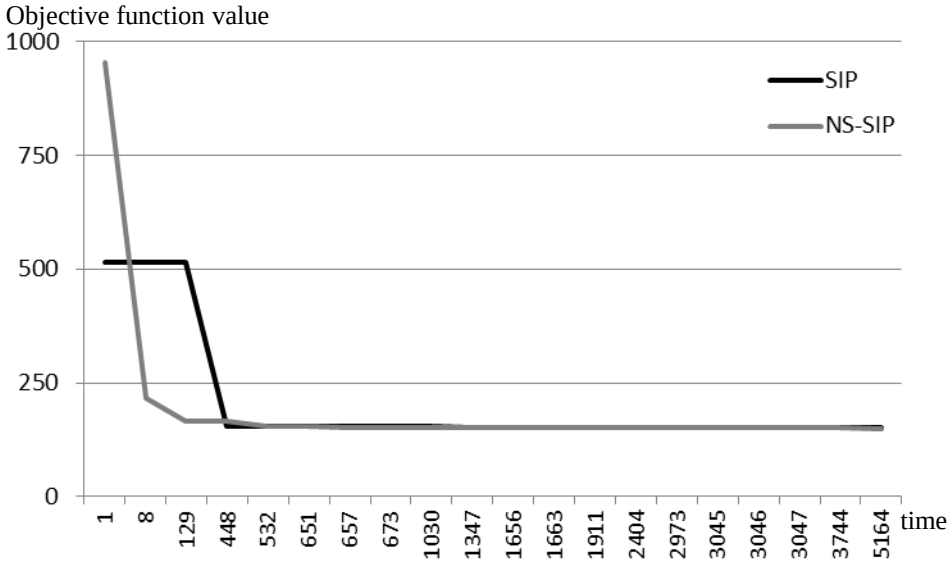


Figure 3.1: Trend of Objective Function Values

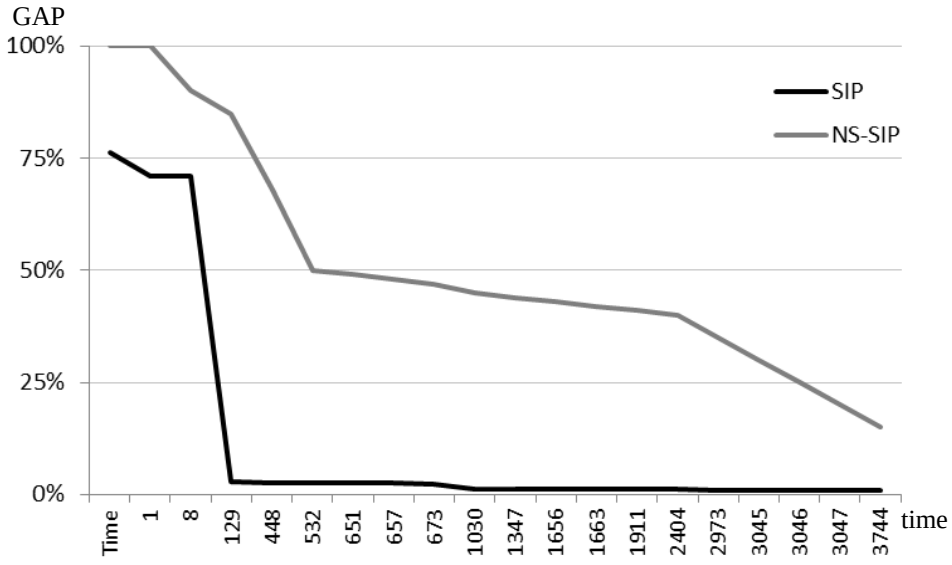


Figure 3.2: Trend of GAP Values

4. MASTER SURGICAL BLOCK SCHEDULES

We deal with the block scheduling policy in this study. A block is the amount of time during which a specific sub-specialty is assigned to an OR. A block may be planned with the duration of two hours, half of a day (e.g., morning, or afternoon), or a day, for example, to permit a surgeon to perform a series of complex surgeries. An alternative, the open scheduling policy, under which each surgeon can schedule his/her surgeries at any time, was common in the 1960s and 1970s but is rarely used in practice today, because it does not utilize surgeons' time as efficiently as block scheduling (Blake et al., 2002).

Table 4.1 illustrates an exemplar block schedule showing that strategic decisions have already assigned the orthopedic specialty to OR-1 and the ophthalmology specialty to OR-2. One sub-specialty of the orthopedic specialty is scheduled in one time block in OR-1, but two sub-specialties of the ophthalmology specialty are scheduled in OR-2, each in one time block.

Table 4.1: An Example of Master Block Surgical Schedule

Time	OR-1	OR-2
8 : 00 ∼ 9 :00	Joint Replacement	Retinal
9 : 00 ∼ 10 :00		
10 : 00 ∼ 11 :00		
11 : 00 ∼ 12 :00		
12 : 00 ∼ 13 :00		
13 : 00 ∼ 14 :00		Pediatric
14 : 00 ∼ 15 :00		
15 : 00 ∼ 16 :00		

MSS determines the duration of the block time (i.e., OR-hours) associated with each sub-specialty and the sequence of time blocks during a day (Beliën and Demuelemeester, 2008), affording each surgeon the opportunity to perform a series

of surgeries efficiently at times acceptable to him/her, while allowing routine office hours. Once MSS determines a schedule of time blocks, including the duration and sequence of each, the day-by-day schedule for a week may be used cyclically, that is, for each week over the intermediate planning horizon. A cyclic schedule avoids the need to prescribe a new schedule every week and promotes coordination among surgeons, staff, and other departments (e.g., PACU, ICU).

The remainder of this paper is organized as follows. Section 4.1 presents preliminaries and section 4.2 describes our solution approach. Section 4.3 presents numerical analysis of the expected values of earliness and lateness associated with normal, lognormal, and gamma surgery durations. Section 4.4 describes the optimal block duration with no-shows. Section 4.5 provides sights for hospital management.

4.1 Preliminaries

This section introduces notation and assumptions used in the subsequent presentation. We also discuss both decision and associated random variables. Finally, we formulate the objective function, which minimizes the sum of expected earliness and lateness costs.

4.1.1 Assumptions and Notation

We deal with a general number of blocks to cast our results in the most generic form possible, even though there may only be one, two, or (at most) four blocks for each OR day. We focus on a single OR because, once surgical specialties have been assigned to OR days, a problem involving multiple ORs can be decomposed into a set of independent problems, each involving a single OR.

We assume that the forecast employed to support the strategic decisions that assign specialties to OR days is compatible with the forecast used at the tactical level to prescribe the MSS, which partitions each OR day into time blocks for subspecialties

associated with the relevant specialty.

We assume that one surgery begins as soon as the previous one ends. Most prior studies have assumed that each surgery is scheduled to begin at the expected completion time of the previous surgery (Choi and Wilhelm, 2012a; Gupta, 2007; Pinedo, 2009), although some incorporate a multiple of the standard deviation of the surgery duration as a safety time to manage risk (Gul et al., 2011). If the previous surgery completes before the scheduled start of the next surgery, OR idleness is incurred; if it finishes after the scheduled start time, the next surgery (i.e., both patient and surgeon who are ready at the scheduled start time) must wait. In contrast, we assume that each surgery begins when the previous surgery ends. This assumption appears to be reasonable in our study because patients are typically prepared well in advance of their scheduled start time and successive surgeries within each block are likely to be performed by the same surgeon so that s/he would be available as well. If successive surgeries are in different time blocks, our assumption would require schedulers to communicate with the surgeon who will perform the next surgery and facilitate her/his readiness ahead of the scheduled start time. This is done currently if possible but may entail establishing different procedures to effect routinely. If the scheduled start time were enforced when the previous surgery is completed early, our assumption would lead to lower bounds on optimal block durations.

Each medical procedure is designated by one of many thousands of CPT codes. Each surgery specialty (e.g., orthopedic) may deal with hundreds of CPT codes and each subspecialty (e.g., joint replacement; bone fractures; knee, spine or shoulder repair) within the index set I of subspecialties that constitute the specialty may deal with dozens of CPT codes. Further, a given surgery may involve a combination of CPT codes. For example, shoulder repair deals with a large number of CPT codes, of which about 15 procedures are performed commonly. Examples of these five-digit

codes are 29805 (diagnostic shoulder arthroscopy), 29826 (shoulder arthroscopy with subacromial decompression), 29807 (labral repair), 29827 (rotator cuff repair), 23430 (bicep tenodesis), and 23120 (acronioclavicular joint resection). One of the authors recently had shoulder-repair surgery that involved the combination of the first four of these CPT codes. We use $\hat{S}_i$ to denote the index set of surgery types associated with subspecialty i, each an individual CPT code or a combination that is common.

We envision a tactical planning process that forecasts n_i, the expected number of surgeries to be performed within subspecialty $i \in I$; and q_{is}, the portion of subspecialty i surgeries that will be of type $s \in \hat{S}_i$. Historical data can be used to estimate $\hat{\mu}_{is}$ and $\hat{\sigma}_{is}$, the mean and variance, respectively, of the duration of surgeries of type $s \in \hat{S}_i$. With this information, the planning process can determine a *representative* surgery duration for each subspecialty, which can be interpreted as the duration of a randomly selected surgery to be performed by this subspecialty. The random duration of the representative surgery, D_i, can be expressed as the convex combination of individual, mutually independent, surgery-type durations $\hat{D}_{is}, s \in \hat{S}_i$: $D_i = \sum_{s \in \hat{S}_i} q_{is} \hat{D}_{is}$. This representative surgery of subspecialty i has a mixture distribution with mean $\mu_i = \sum_{s \in \hat{S}_i} q_{is} \hat{\mu}_{is}$ and variance $\sigma_i^2 = \sum_{s \in \hat{S}_i} q_{is}^2 \hat{\sigma}_{is}^2$ and, by the Central Limit Theorem (CLT) (Casella and Berger, 2001), is normally distributed because $|\hat{S}_i|$, for each $i \in I$, is large, as described in the paragraph above. Based on this analysis, we treat the duration of surgeries of subspecialty i as *i.i.d.* normal random values.

We now define the notation we use in the subsequent presentation. If subspecialty i is assigned to one block and the duration of each surgery is D_i hours with mean μ_i and variance σ_i^2, the block must accommodate the total surgery time, the n_i-fold convolution of D_i, which has mean $\bar{\mu}_i := n_i \mu_i$ and variance $\bar{\sigma}_i^2 := n_i \sigma_i^2$.

We use map $\Delta : I \to K$ to represent a set of sequences (or permutations), each

Index Sets and Indices

I	sub-specialties	$i \in I$
K	sequence positions for time blocks	$k \in K$
Δ	permutations of time blocks	$\delta \in \Delta$

Parameters

c^e	Earliness penalty cost	
c^l	Lateness penalty cost	
β	Ratio of earliness cost to lateness cost,	$\beta = c^e/c^l$.

of which assigns each sub-specialty to one and only one sequence position. We use subscripts $[k]$ for kth block sequence position and i for sub-specialty to avoid potential confusion. The total number of permutations is $|K|!$, where $|K|$ is the number of blocks and $|K| = |I|$.

Decision variables prescribe planned block durations (i.e., $x^{\delta}_{[k]}$ gives the planned duration of the kth block) and block sequence (i.e., permutation δ) for one day in the OR. The planned end time of the block in the kth position, given δ, is prescribed by $y^{\delta}_{[k]}$, where $y^{\delta}_{[k]} = x^{\delta}_{[1]} + \cdots + x^{\delta}_{[k]}$. The planned end time of the last block $y^{\delta}_{[|K|]}$ corresponds to the end of the OR day and is important in deciding the number of hours that the staff will be required to work and the amount of overtime that is required. The utilization of an OR, as determined by $y^{\delta}_{[|K|]}$, relative to the length of the work day may vary by day because it depends on forecast workloads and sub-specialties assigned to each day.

The random duration of the kth block in the sequence, $B^{\delta}_{[k]}$ is the $n_{[k]}$-fold convolution of $D_{[k]}$ and has mean $\bar{\mu}_{[k]} = n_{[k]}\mu_{[k]}$ and variance $\bar{\sigma}^2_{[k]} = n_{[k]}\sigma^2_{[k]}$. Random duration $B^{\delta}_{[k]}$ must be compared with decision variable $x^{\delta}_{[k]}$, which prescribes planned block duration. We define $T^{\delta}_{[k]} := B^{\delta}_{[1]} + \cdots + B^{\delta}_{[k]}$ as the random end time to complete all surgeries assigned to blocks [1] through $[k]$ and compare it with decision variable

$y^{\delta}_{[k]}$, the planned end time of block $[k]$.

Figure 4.1 shows the relationship between decision variables and related random variables for four time blocks: the former are indicated below the time line; and the latter, above. In Figure 4.1, lateness is incurred in association with the third block; earliness, with other blocks. If the last surgery of the sub-specialty assigned to the block completes earlier than the planned completion time of the block, earliness results; otherwise, lateness is incurred. Because no surgery will be started after the last block, its earliness corresponds to surgeon and OR idleness. The lateness of the last block corresponds to overtime.

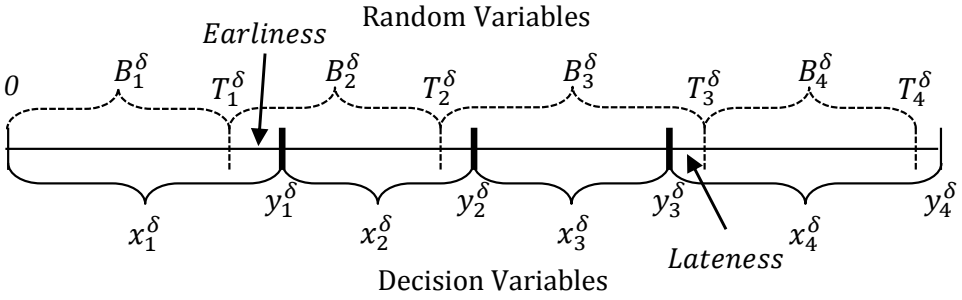


Figure 4.1: Relationship between Decision and Random Variables

Our objective function penalizes the expected earliness, $E[(x^{\delta}_{[1]}+\cdots+x^{\delta}_{[k]}-B^{\delta}_{[1]}-\cdots-B^{\delta}_{[k]})^{+}]$ or $E[(y^{\delta}_{[k]}-T^{\delta}_{[k]})^{+}]$ and the expected lateness, $E[(B^{\delta}_{[1]}+\cdots+B^{\delta}_{[k]}-x^{\delta}_{[1]}-\cdots-x^{\delta}_{[k]})^{+}]$ or $E[(T^{\delta}_{[k]}-y^{\delta}_{[k]})^{+}]$, of each block $k \in K$. The former represents the cost of expediting the start time of the next surgery; and the latter, the cost of delaying the start time of the next surgery. In the case of the last surgery in the sequence, the cost of earliness represents idleness, and the cost lateness represents overtime premium. We use the same lateness penalty, c^{l}, for all blocks $k \in K$ because analytical results depend mainly upon parameters of surgery durations (i.e., mean and standard deviation). On-time performance is important in health-care

delivery systems, because MSS coordinates surgeons, nurses, and anesthesiologists and influences other departments like PACU. We build a schedule that balances the expected costs of earliness and lateness associated with each block, defining objective functions $f_{[k]}(x^{\delta}_{[1]} + \cdots + x^{\delta}_{[k]})$ and $f_{[k]}(y^{\delta}_{[k]}), k \in K, \delta \in \Delta$, respectively, as follows:

$$\begin{aligned} f_{[k]}(x^{\delta}_{[1]} + \cdots + x^{\delta}_{[k]}) := & c^e E\Big[\big((x^{\delta}_{[1]} + \cdots + x^{\delta}_{[k]} - B^{\delta}_{[1]} - \cdots - B^{\delta}_{[k]}\big)^+\Big] \\ & + c^l E\Big[\big(B^{\delta}_{[1]} + \cdots + B^{\delta}_{[k]} - x^{\delta}_{[1]} - \cdots - x^{\delta}_{[k]}\big)^+\Big], \qquad (4.1) \\ f_{[k]}(y^{\delta}_{[k]}) := & c^e E[(y^{\delta}_{[k]} - T^{\delta}_{[k]})^+] + c^l E[(T^{\delta}_{[k]} - y^{\delta}_{[k]})^+]. \qquad (4.2) \end{aligned}$$

4.1.2 Mathematical Model

This subsection describes two optimization models, one in terms of $x^{\delta}_{[k]}$ and another, which transforms $x^{\delta}_{[k]}$ to decision variable $y^{\delta}_{[k]}$. We focus on the latter model to prescribe optimal durations and sequence, because it reduces a complicated problem to a series of newsvendor problems, the *sequential newsvendor* model. In this subsection, we describe the two mathematical models and show how to exploit the *sequential newsvendor* model. Lastly, we depict the *sequential newsvendor* model graphically.

In minimizing the sum of expected earliness and lateness costs, the *sequential newsvendor problem (SNV)*, in terms of decision variable $x^{\delta}_{[k]}, k \in K$, is :

$$\begin{aligned} (SNV^{\delta}(x)) \qquad & \min_{\delta \in \Delta} \min_{x^{\delta}_{[k]}: k \in K} \sum_{k \in K} f_{[k]}(x^{\delta}_{[1]} + \cdots + x^{\delta}_{[k]}) \qquad (4.3) \\ & s.t. \qquad x^{\delta}_{[k]} \geq 0 \qquad k \in K, \delta \in \Delta. \qquad (4.4) \end{aligned}$$

We seek to determine the optimal planned duration $\hat{x}^{\hat{\delta}}_{[k]}$ of the kth block, $k \in K$ and the optimal block sequence (i.e., permutation) $\hat{\delta}$. To solve $SNV^{\delta}(x)$, we need to show that both the first order necessary condition (FONC) and the second order necessary condition (SONC) are satisfied; i.e., the Hessian matrix of the objective function should be semi-positive at the optimal point $\hat{x}^{\hat{\delta}}_{[k]}$ (Bazaraa et al., 2006). Constraint (4.4) requires decision variable $x^{\delta}_{[k]}, k \in K$ to be non-negative. Solving the problem with decision variables $x^{\delta}_{[k]}, k \in K$ is challenging because it requires a complex Hessian matrix to be evaluated.

Instead, we employ a linear transformation to use alternative decision variables $y_{[k]}, k \in K$ and do not use a Hessian matrix, as described in the following subsection. With planned end-time decision variable $y_{[k]}, k \in K$, problem $SNV^{\delta}(x)$ can be transformed to $SNV^{\delta}(y)$:

$$(SNV^{\delta}(y)) \qquad \min_{\delta \in \Delta} \min_{y^{\delta}_{[k]}: k \in K} \sum_{k \in K} f_{[k]}(y^{\delta}_{[k]}) \tag{4.5}$$

$$s.t. \qquad y^{\delta}_{[k-1]} \leq y^{\delta}_{[k]} \qquad k = 2, \ldots, |K|, \delta \in \Delta \tag{4.6}$$

$$y^{\delta}_{[k]} \geq 0 \qquad k \in K, \delta \in \Delta. \tag{4.7}$$

Figure 4.2 depicts the variable transformation in two dimensional space. While the feasible area is the first quadrant of x-space, owing to constraints (4.4), the feasible area in y-space is half of the first quadrant as shown in Figure 4.2(b) owing to constraints (4.6) and (4.7).

The variable transformation from x-space to y-space recasts $SNV^{\delta}(x)$ as $SNV^{\delta}(y)$, which is able to utilize well-known properties of the newsvendor problem. Given δ,

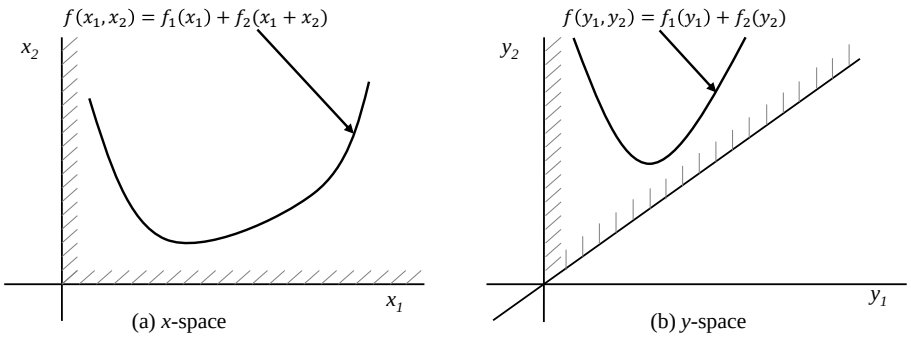


Figure 4.2: Variable Transformation from x-space to y-space

each $[k]$ term in (4.5),

$$(NV^{\delta}_{[k]}(y^{\delta}_{[k]})) \qquad \min_{y^{\delta}_{[k]}} f_{[k]}(y^{\delta}_{[k]}),$$

is a newsvendor problem. $f_{[k]}(y^{\delta}_{[k]})$, equivalently $f_{[k]}(x^{\delta}_{[1]} + \cdots + x^{\delta}_{[k]})$, is a convex objective function and its solution satisfies FONC and SONC at the optimal point (Porteus, 2002). We discuss constraint (4.6), which requires $y^{\delta}_{[k]}$ to increase with k, in the following section. Proposition 1 establishes that $f_{[k]}(y_{[k]})$ is convex. The sum of $f_{[k]}(y_{[k]})$'s is a convex function because the sum of convex functions is also convex (Bazaraa et al., 2006). Hence, the objective functions of both $SNV^{\delta}(x)$ and $SNV^{\delta}(y)$ are convex. Our development exploits these simple characteristics of the newsvendor problem. We now suppress superscript δ to streamline presentation.

Proposition 1. *Because $f_{[k]}(y_{[k]})$, equivalently $f_{[k]}(x_{[1]} + \cdots + x_{[k]})$, is convex, it satisfies FONC and SONC at the optimal point, say $\hat{y}_{[k]}, k \in K$. The optimal solution to $NV_{[k]}(y_{[k]}), \hat{y}_{[k]}$, is the value at which the distribution function $F_{T_{[k]}}(y_{[k]})$ is equal to the critical ratio:*

$$F_{T_{[k]}}(\hat{y}_{[k]}) = \frac{c^l}{c^e + c^l} = \frac{1}{1+\beta}. \tag{4.8}$$

Critical ratio $\frac{1}{1+\beta}$ *is the same for each* $k \in K$ *because cost parameters are the same for all blocks.*

Proof. Because $f_{[k]}(y_{[k]}), k \in K$ is the objective function of a newsvendor-type problem that balances under- and over-age, it is known to be convex (Porteus, 2002). Hence, the equivalent function $f_{[k]}(x_{[1]} + \cdots + x_{[k]})$ is also convex. The newsvendor objective function $f_{[k]}(y_{[k]})$ satisfies FONC and SONC (Porteus, 2002) at optimal point $\hat{y}_{[k]}$ so that

$$f'_{[k]}(\hat{y}_{[k]}) = 0 \text{ and } f''_{[k]}(\hat{y}_{[k]}) \geq 0, k \in K.$$

The solution $\hat{y}_{[k]}$ specified by (4.8) is known to optimize $NV_{[k]}(y_{[k]})$ (Porteus, 2002; Nahmias, 2008). □

We now introduce a new function, $g(\delta)$, to explain the *sequential newsvendor* problem:

$$Z^* = \min_{\delta \in \Delta}\{g(\delta) : (4.6), (4.7), \text{ and } g(\delta) = \min \sum_{k \in K} f_{[k]}(y^{\delta}_{[k]})\}. \tag{4.9}$$

Figure 4.3 depicts the *sequential newsvendor* problem, representing relationships among $g(\delta)$ and $NV^{\delta}_{[k]}(y^{\delta}_{[k]})$. There are $|\Delta| = |K|!$ possible sequences, and each $\delta \in \Delta$ has an associated $g(\delta)$. Z^* of (4.9) achieves its minimum at sequence $\hat{\delta}$.

Given sequence $\delta \in \Delta$, $NV^{\delta}_{[k]}(y^{\delta}_{[k]})$ defines the objective function of a newsvendor problem that prescribes the optimal, planned end time of block $[k]$, $\hat{y}^{\delta}_{[k]}, k \in K$. Given δ, $g(\delta)$ is the sum of optimal values to $NV^{\delta}_{[k]}(y^{\delta}_{[k]})$, for all $k \in K$. Z^* in (4.9) is minimized by prescribing the best sequence (i.e., permutation) of blocks. We first determine $g(\delta)$ by summing the solutions of the $|K|$ newsvendor problems, giving the optimal block durations for a given sequence δ, then find the best sequence as described in following section.

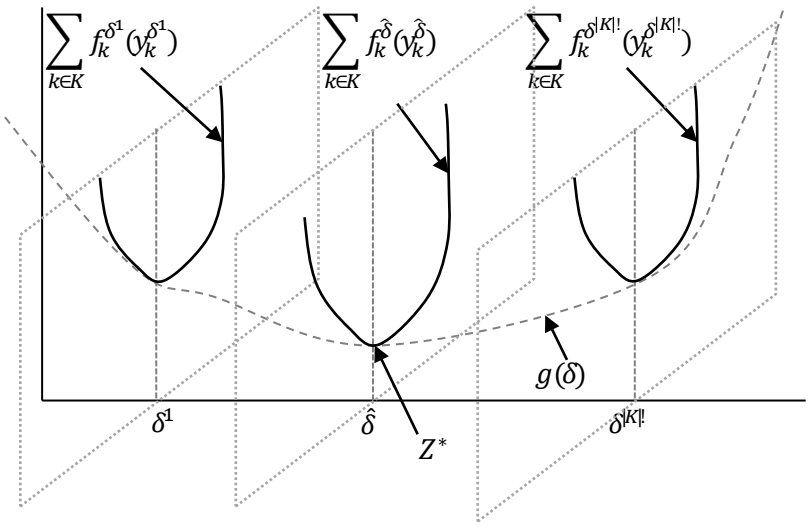


Figure 4.3: Sequential Newsvendor Problem

4.2 Solution Approach

In this section, we describe solution approaches to prescribe optimal block durations and optimal sequence. We show that the newsvendor solution, $NV^{\delta}_{[k]}(y^{\delta}_{[k]}), k \in K$, gives the optimal planned end times for a given sequence $\delta \in \Delta$. We derive the closed form of the objective value and prove that the SV rule is optimal to sequence blocks for the case in which surgeries are independent and normally distributed.

Subsection 4.2.1 devises optimal block durations for the unconstrained version of $SNV(y)$ and subsection 4.2.2 devises the closed-form solution to the constrained version. Subsection 4.2.3 determines the optimal sequence, $\hat{\delta}$.

4.2.1 Unconstrained Optimal Block Durations

We first seek the unconstrained (i.e., without constraints (4.6) and (4.7)) optimal block durations for a given sequence δ . In this subsection, we assume that sequence δ is fixed, so we suppress this superscript. We may solve problem $SNV(x)$ using a

dynamic programming approach; however, it is hard to prove optimality for a general number of blocks, because $x_{[k]}$ appears in $f_{[k]}(x_{[1]}+\cdots+x_{[k]}), \ldots,$ and $f_{[|K|]}(x_{[1]}+\cdots+x_{[|K|]})$ so that these functions are not separable. However, Proposition 2 establishes separability of the transformed problem, $SNV(y)$. Subsequently, we are able to solve independent newsvendor problems $NV_{[k]}(y_{[k]}), k \in K$, owing to separability.

Proposition 2. *Define the unconstrained sequential newsvendor problem* $USNV(y)$ *by relaxing constraints* (4.6) *and* (4.7) *to obtain:*

$$(USNV(y)) \qquad \min \sum_{k \in K} f_{[k]}(y_{[k]}). \tag{4.10}$$

Problem $USNV(y)$ *is separable with respect to* $y_{[k]}$*:*

$$\begin{aligned} & \min \left\{ f_{[1]}(y_{[1]}) + f_{[2]}(y_{[2]}) + \cdots + f_{[|K|]}(y_{[|K|]}) \right\} \\ \equiv \quad & \min f_{[1]}(y_{[1]}) + \min f_{[2]}(y_{[2]}) + \cdots + \min f_{[|K|]}(y_{[|K|]}). \end{aligned} \tag{4.11}$$

Proof. After relaxing (4.6) and (4.7), problem $USNV(y)$ is separable with respect to $y_{[k]}, k \in K$ because $NV_{[k]}(y_{[k]}), k \in K$ is independent of other variables $y_{[k']}, k'(\neq k) \in K$. □

Based on Proposition 2, we can solve individual $NV_{[k]}(y_{[k]}), k \in K$ problems independently to optimize $USNV(y)$; any $\hat{y}_{[k]}, k \in K$, that satisfies FONC and SONC optimizes $USNV(y)$. $NV_{[k]}(y_{[k]})$ is a newsvendor-type problem that prescribes the planned end time of the $[k]$th block to minimize the sum of expected earliness and lateness costs. Newsvendor problem $NV_{[k]}(y_{[k]})$, which is associated with random variable $T_{[k]}, k \in K$, can be solved independently according to Proposition 3, which follows. Even though $T_{[k]}$'s are not independent random variables, we can solve problems $NV_{[k]}(y_{[k]}), k \in K$ independently after relaxing constraints (4.6) because

$E[(y_{[k]} - T_{[k]})^+]$ and $E[(T_{[k]} - y_{[k]})^+]$ are functions of $y_{[k]}$ and $T_{[k]}$ is essentially a parameter that gives information about all surgeries through the kth block.

Proposition 3. *Given random block duration $B_{[k]}$ with mean $\bar{\mu}_k$ and variance $\bar{\sigma}^2_{[k]}, k \in K$, let $T_{[k]}$ be the sum of the independent, normally distributed random durations of surgeries associated with sub-specialties, each assigned to a block [1] through [k] (i.e., $T_{[k]} := B_{[1]} + \cdots + B_{[k]}$) and let $F_{T_{[k]}}$ be the normal distribution function of $T_{[k]}$, which has mean $\bar{\bar{\mu}}_{[k]} := \bar{\mu}_{[1]} + \cdots + \bar{\mu}_{[k]}$ and variance $\bar{\bar{\sigma}}^2_{[k]} := \bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k]}$.*

(i) *Problem $NV_{[k]}(y)$, which prescribes the optimal planned end time of the kth block, has optimal solution, $\hat{y}_{[k]}$ for each $k \in K$ such that:*

$$F_{T_{[k]}}(\hat{y}_{[k]}) \quad = \quad \frac{c^l}{c^e + c^l} = \frac{1}{1+\beta} = \Phi(z), \tag{4.12}$$

where $T_{[k]} \sim N(\bar{\bar{\mu}}_{[k]}, \bar{\bar{\sigma}}^2_{[k]})$; $\hat{y}_{[k]} = \bar{\bar{\mu}}_{[k]} + z\bar{\bar{\sigma}}_{[k]}, k \in K$; $\Phi(z)$ is the standard normal distribution function; and z is the normal score.

(ii) *The corresponding, optimal block duration, $\hat{x}_{[k]}$ for each $k \in K$, can be obtained by definition: $\hat{x}_{[1]} = \hat{y}_{[1]}$, and $\hat{x}_{[k]} = \hat{y}_{[k]} - \hat{y}_{[k-1]}, k = 2, \ldots, |K|$.*

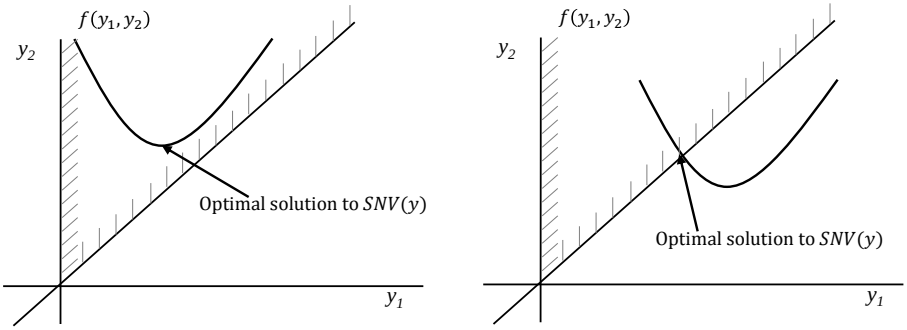
Proof. (i) $T_{[k]} = B_{[1]} + \cdots + B_{[k]}$, $\bar{\bar{\mu}}_k = E[T_{[k]}] = E[B_{[1]}] + \cdots + E[B_{[k]}] = \bar{\mu}_1 + \cdots + \bar{\mu}_k$ and $\bar{\bar{\sigma}}^2_{[k]} = V[T_{[k]}] = V[B_{[1]}] + \cdots + V[B_{[k]}] = \bar{\sigma}^2_1 + \cdots + \bar{\sigma}^2_k, k \in K$. Because individual $B_{[k]}$ are normally distributed, $T_{[k]}$ is also normally distributed. $\hat{y}_{[k]}$, as defined by (4.12), is the optimal solution to newsvendor problem $NV_{[k]}(y_{[k]})$ (Nahmias, 2008; Porteus, 2002).

(ii) follows from the definition of the variable transformation from x to y. □

4.2.2 Constrained Optimal Block Durations

Now, we solve constrained optimization problem $SNV(y)$, focusing on constraints (4.6) and (4.7), which require $0 \leq y_{[k-1]} \leq y_{[k]}, k = 2, \ldots, |K|$, correspondingly, that

$0 \leq x_{[k]}, k \in K$. According to the Karush Kuhn Tucker (KKT) conditions, if the optimal solution to unconstrained problem $USNV(y)$ satisfies constraints (4.6) and (4.7), it is the global optimal solution to constrained problem $SNV(y)$ as shown Figure 4.4(a). Otherwise, the optimal solution is on the boundary so that $\hat{y}^{\delta}_{[k-1]} = \hat{y}^{\delta}_{[k]}$ for one or more $k \in K$ as shown Figure 4.4(b).



(a) Optimal solutions to SNV(y) and USNV(y) are the same.

(b) Optimal solutions to SNV(y) is on the boundary.

Figure 4.4: Graphical Depiction of KKT Conditions

We show the KKT conditions analytically. At the constrained optimal solution

$\hat{y}_{[k]}$ to $SNV(y)$, KKT conditions must hold (Bazaraa et al., 2006):

$$\begin{pmatrix} \frac{\partial}{\partial y_{[1]}} f_{[1]}(\hat{y}_{[1]}) & 0 & \dots & \dots & 0 \\ 0 & \frac{\partial}{\partial y_{[2]}} f_{[2]}(\hat{y}_{[2]}) & 0 & \dots & 0 \\ 0 & \dots & \frac{\partial}{\partial y_{[k]}} f_{[k]}(\hat{y}_{[k]}) & \dots & 0 \\ 0 & \dots & \dots & 0 & \frac{\partial}{\partial y_{[|K|]}} f_{[|K|]}(\hat{y}_{[|K|]}) \end{pmatrix}$$

$$+u_{[1]} \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & -1 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 0 \end{pmatrix} + u_{[2]} \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & -1 & \dots & 0 \\ 0 & \dots & \dots & 0 & 0 \end{pmatrix} + \cdots$$

$$+u_{[|K|]} \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & \dots & 0 & 0 \\ 0 & \dots & \dots & 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix} \tag{4.13}$$

$$u_{[k-1]}(\hat{y}_{[k-1]} - \hat{y}_{[k]}) = 0 \qquad k = 2, \dots, |K|, \tag{4.14}$$

where $u_{[k]}$ is a Lagrangian multiplier associated with constraint (4.6) of $k \in K$. If the optimal solution to $USNV(y)$ is feasible with respect to (4.6) and (4.7) as shown in Figure 4.4 (a), $\frac{\partial}{\partial y_{[k]}} f_{[k]}(\hat{y}_{[k]}) = 0$ and $u_{[k]} = 0$ hold for all $k \in K$. Because $u_{[k]} = 0$, $\hat{y}_{[k]}$ is not necessarily equal to $\hat{y}_{[k+1]}$. If the optimal solution to unconstrained problem $USNV(y)$ violates either constraints (4.6) or (4.7) as shown Figure 4.4 (b), $\frac{\partial}{\partial y_{[k]}} f_{[k]}(\hat{y}_{[k]}) \neq 0$ and the optimal solution lies on the boundary, so that $\hat{y}_{[k-1]} = \hat{y}_{[k]}$ (i.e., $\hat{x}_{[k]} = 0$) and the Lagrangian multiplier $u_{[k]}$ is non-zero for some $k \in K$ (Bazaraa et al., 2006). In our analysis, we concentrate on the former case, which is depicted by Figure 4.4 (a).

Next, we derive a condition to assure that a solution to unconstrained problem $USNV(y)$ satisfies (4.6) and (4.7). Even though random variable $T_{[k]}$ has a larger mean and variance than $T_{[k-1]}$, it is numerically possible, but not practically feasible, for an optimal solution to violate (4.6) (i.e., $\hat{y}_{[k]} < \hat{y}_{[k-1]}$ for some k). We first give an example to demonstrate the relevant issues. Consider two normal distributions representing block surgery durations with $\mu_1 = 2, \sigma_1 = 0.1$ and $\mu_2 = 1, \sigma_2 = 0.7$. Then, for sequence $1 \rightarrow 2$, $\mu_{[1]} = 2, \mu_{[2]} = 3, \sigma_{[1]} = 0.1$ and $\sigma_{[2]} = \sqrt{.1^2 + .7^2}$. Assume that $\beta = 25$. Optimal solutions $\hat{y}_{[1]}$ and $\hat{y}_{[2]}$ are such that

$$F_{T_{[1]}}(y_{[1]}) = F_{T_{[2]}}(y_{[2]}) = \frac{1}{1+25} = 0.038, \tag{4.15}$$

so that $z = -1.768$. Then $\hat{y}_{[1]} = \mu_{[1]} - 1.768\sigma_{[1]} = 1.823$ and $\hat{y}_{[2]} = \mu_{[2]} - 1.768\sigma_{[2]} = 1.749$, so that $y_{[1]} > y_{[2]}$ and these values are not feasible with respect to (4.6). In this case, the planned end time of the second block is less than the planned end time of the first block, which would mean that the planned duration of the second block were negative, i.e., $\hat{x}_{[1]} = 1.823$ and $\hat{x}_{[2]} = -0.074$. This example is an extreme case because $T_{[2]}$ has a much larger variance and a smaller mean than $T_{[1]}$. Furthermore, the ratio of the two costs, β, is huge because the cost of earliness is 25 times of the cost of lateness.

Considering two consecutive blocks, Proposition 4 establishes restrictions on parameters to assure that mathematically feasible, optimal solutions are also practically feasible. Based on the definition of $y_{[k]} = \bar{\bar{\mu}}_{[k]} + z\bar{\bar{\sigma}}_{[k]}$, we impose condition $\bar{\mu}_k \geq |z|\bar{\sigma}_k, k \in K$, which, in turn, assures that $y_{[k]} \geq 0$ holds. Proposition 4 establishes that this condition also implies that $y_{[1]} \leq y_{[2]}$ relative to two independent and normally distributed distributions and this result is extended to all $k \in K$ by Corollary 5.

Proposition 4. *Consider two block durations B_i, which are independent and normally distributed: $N(\bar{\mu}_i, \bar{\sigma}_i^2), i = 1, 2$. We require that $F_{B_1}(y_{[1]}) = F_{B_1+B_2}(y_{[2]}) = 1/(1+\beta) = \Phi(z)$ so that $y_{[1]} = \bar{\mu}_1 + z\bar{\sigma}_1$ and $y_{[2]} = \bar{\mu}_1 + \bar{\mu}_2 + z\sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2}$. If $\bar{\mu}_i \geq |z|\bar{\sigma}_i, i = 1, 2$, then $y_{[1]} \leq y_{[2]}$; in other words,*

$$\bar{\mu}_1 + z\bar{\sigma}_1 \leq \bar{\mu}_1 + \bar{\mu}_2 + z\sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2}. \tag{4.16}$$

Proof. Case (i): $z \geq 0$. This case occurs if $\beta \leq 1$. Using $+|z|$ to denote $z \geq 0$, inequality (4.16) becomes

$$\bar{\mu}_1 + |z|\bar{\sigma}_1 \leq \bar{\mu}_1 + \bar{\mu}_2 + |z|\sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2}. \tag{4.17}$$

In this case, equation(4.17) is trivially true.

Case (ii): $z < 0$. This case occurs if $\beta > 1$. Using $-|z|$ to denote $z < 0$, the equivalent of inequality (4.16) is

$$\bar{\mu}_1 - |z|\bar{\sigma}_1 \leq \bar{\mu}_1 + \bar{\mu}_2 - |z|\sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2}. \tag{4.18}$$

We must now show that (4.16) holds in the form of (4.18) in case (ii).

Combining the fundamental relationship $\sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2} \leq \bar{\sigma}_1 + \bar{\sigma}_2$ with conditions, $\bar{\mu}_i \geq |z|\bar{\sigma}_i, i = 1, 2$, the following inequality holds:

$$|z|(\sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2} - \bar{\sigma}_1) \leq |z|\bar{\sigma}_2 \leq \bar{\mu}_2.$$

Adding $\bar{\mu}_1$ to the left- and right-most terms,

$$\bar{\mu}_1 + |z|(\sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2} - \bar{\sigma}_1) \leq \bar{\mu}_1 + \bar{\mu}_2,$$

which can be rearranged to establish (4.18) when $z < 0$, showing that inequality (4.16) holds for both positive and negative z values. □

We now generalize Proposition 3 to prescribe planned end times for all blocks, relying on the Proposition 4, which deals with a two-block case.

Corollary 5. *Given* $B_{[k]} \sim N(\bar{\mu}_{[k]}, \bar{\sigma}^2_{[k]})$ *such that* $\bar{\mu}_{[k]} \geq |z|\bar{\sigma}_{[k]}, k \in K$, *let* $T_{[k]} := B_{[1]} + \cdots + B_{[k]}$ *have mean* $\bar{\bar{\mu}}_{[k]} = \bar{\mu}_{[1]} + \cdots + \bar{\mu}_{[k]}$ *and variance* $\bar{\bar{\sigma}}^2_{[k]} = \bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k]}$. *Optimal block durations* $\hat{x}_{[k]}, k \in K$ *can be obtained from optimal planned end times* $\hat{y}_{[k]}, k \in K$ *as follows:*

$$\hat{x}_{[1]} = \hat{y}_{[1]} \qquad = \bar{\bar{\mu}}_{[1]} + z\bar{\bar{\sigma}}_{[1]} = \bar{\mu}_{[1]} + z\bar{\sigma}_{[1]} \tag{4.19}$$

$$\begin{aligned} \hat{x}_{[k]} = \hat{y}_{[k]} - \hat{y}_{[k-1]} &= \bar{\bar{\mu}}_{[k]} + z\bar{\bar{\sigma}}_{[k]} - \bar{\bar{\mu}}_{[k-1]} - z\bar{\bar{\sigma}}_{[k-1]} \\ &= \bar{\mu}_{[k]} + z\left[\sqrt{\sum_{l=1}^{k} \bar{\sigma}^2_{[l]}} - \sqrt{\sum_{l=1}^{k-1} \bar{\sigma}^2_{[l]}}\right], \quad k \geq 2 \end{aligned} \tag{4.20}$$

Proof. If condition $\bar{\mu}_{[k]} \geq |z|\bar{\sigma}_{[k]}, k \in K$ of Proposition 4 is satisfied, $\hat{y}_{[k-1]} \leq \hat{y}_{[k]}, k = 2, \ldots, |K|$, because Proposition 4 can be applied to each pair of successive blocks (e.g., [1] and [2], [2] and [3], and so on). The proof relies on the fact that each $T_{[k-1]}$ is normally distributed (i.e., equivalent to B_1 in Proposition 4) and each $B_{[k]}$ is independent and normally distributed so that $T_{[k]} = T_{[k-1]} + B_{[k]}$ (i.e., equivalent to $B_1 + B_2$ in Proposition 4), where $B_1 \sim N(\bar{\mu}_1 + \cdots + \bar{\mu}_{[k-1]}, \bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k-1]})$ and $B_2 \sim N(\bar{\mu}_{[k]}, \bar{\sigma}^2_{[k]})$. The optimal planned end time $\hat{y}_{[k]}$ of the kth block is given by

$$\hat{y}_{[k]} = \sum_{l=1}^{k} \bar{\mu}_{[l]} + z\sqrt{\sum_{l=1}^{k} \bar{\sigma}^2_{[l]}}.$$

Optimal block duration $\hat{x}_{[k]}$ can be obtained by definition: $\hat{x}_1 = \hat{y}_1 \geq 0$ and $\hat{x}_{[k]} = \hat{y}_{[k]} - \hat{y}_{[k-1]} \geq 0, k \geq 2$. □

If condition $\bar{\mu}_{[k]} \geq |z|\bar{\sigma}_{[k]}, k \in K$ is satisfied, $\bar{\bar{\mu}}_{[k]} \geq |z|\bar{\bar{\sigma}}_{[k]}, k \in K$ is also satisfied because $\bar{\mu}_{[1]} + \cdots + \bar{\mu}_{[k]} \geq |z|(\bar{\sigma}_{[1]} + \cdots + \bar{\sigma}_{[k]}) \geq |z|\sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k]}}$. If condition $\bar{\mu}_{[k]} \geq |z|\bar{\sigma}_{[k]}, k \in K$ is satisfied, both optimal planned end-times $\hat{y}_{[k]}, k \in K$ and optimal block durations $\hat{x}_{[k]}, k \in K$ will be non-negative; accordingly, $0 \leq \hat{y}_{[k-1]} \leq \hat{y}_{[k]}, k = 2, \ldots, |K|$. If we use different cost parameter values for the last block to reflect the fact that lateness for this block is actually overtime and earliness is idleness, optimal solutions, $\hat{x}_{[|K|]}, k \in K$ are given as follows:

$$\hat{x}_{[|K|]} = \hat{y}_{[|K|]} - \hat{y}_{[|K|-1]} = \bar{\mu}_{[|K|]} + \bar{z}\sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[|K|]}} - z\sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[|K|-1]}}, \quad (4.21)$$

such that $\Phi(\bar{z}) = \frac{\bar{c}^l}{\bar{c}^l + \bar{c}^e} = \frac{1}{1+\bar{\beta}}$, where $\bar{c}^l$ corresponds to lateness (i.e, overtime) cost; $\bar{c}^e$, to earliness (i.e., idleness) cost. The optimal solutions depend mainly upon parameters such as mean and variance, if two cost ratios (i.e., β and $\bar{\beta}$) are not significantly different.

4.2.3 Optimal Block Sequence

$NV_{[k]}(y_{[k]})$ defines a newsvendor problem that prescribes the planned end time of block $[k]$; $SNV(y)$ seeks the sum of optimal solutions to all $NV_{[k]}(y_{[k]}), k \in K$ and defines the $g(\delta)$ value for each permutation δ. Z^* in (4.9) is the objective function value associated with the optimal sequence of newsvendor solutions; i.e., the minimum of $g(\delta)$ over sequences $\delta \in \Delta$. Hence, the next problem we solve is to determine the optimal sequence, $\hat{\delta}$, which we address as the *sequential newsvendor* problem. In the previous subsection, we prescribe optimal block durations under a fixed sequence. We want to find the minimum $g(\delta)$ over all $\delta \in \Delta$.

We show that each $g(\delta)$ can be expressed in a closed form when surgery durations assigned to each block are independent and normally distributed and use this form to derive the optimal rule to sequence blocks. Consider the duration of the kth block, $y_{[k]}$. We suppress superscripts and subscripts for clarity, defining $f(y)$ and $T \sim N(\bar{\mu}, \sigma^2)$, so that the objective function of $NV_{[k]}(y_{[k]})$ becomes

$$f(y) = c^e E[(y - T)^+] + c^l E[(T - y)^+].$$

We use Lemmas 6 - 8 to derive a closed form expression for the optimal value of $\min f(y)$.

Lemma 6. *For $T \sim N(\bar{\bar{\mu}}, \bar{\bar{\sigma}}^2)$,*

$$E[(y - T)^+] = \frac{\bar{\bar{\sigma}}}{\sqrt{2\pi}} e^{-\frac{(y-\bar{\bar{\mu}})^2}{\bar{\bar{\sigma}}^2}} + (y - \bar{\bar{\mu}})\Phi(\frac{y - \bar{\bar{\mu}}}{\bar{\bar{\sigma}}}).$$

Proof. See the Appendix. □

Lemma 7. *For $T \sim N(\bar{\bar{\mu}}, \bar{\bar{\sigma}}^2)$,*

$$E[(T - y)^+] = \frac{\bar{\bar{\sigma}}}{\sqrt{2\pi}} e^{-\frac{(y-\bar{\bar{\mu}})^2}{\bar{\bar{\sigma}}^2}} + (\bar{\bar{\mu}} - y)(1 - \Phi(\frac{y - \bar{\bar{\mu}}}{\bar{\bar{\sigma}}})).$$

Proof. See the Appendix. □

Now, we simplify $Z = \min_y f(y)$, expressing Z as an increasing function of $\bar{\bar{\sigma}}$. We invoke Lemma 8 for a single block.

Lemma 8. *If $TN(\bar{\bar{\mu}}, \bar{\bar{\sigma}}^2)$, the optimal value $\hat{Z}$ of the problem $\min_y f(y)$ is defined as:*

$$\hat{Z} = (c^e + c^l)\frac{\bar{\bar{\sigma}}}{\sqrt{2\pi}} e^{-z^2},$$

where $\Phi(z) = \frac{1}{1+\beta} = \frac{c^l}{c^e+c^l}$.

Proof. See the Appendix. □

We next apply Lemma 8 to a particular sequence to obtain a closed-form for $g(\delta)$ for a general number of blocks.

Proposition 9. *Objective function* $g(\delta^1)$*, evaluated for a particular sequence, say* $\delta^1 : 1 \to 2 \to \cdot \to |K|$ *(i.e.,* $[k] = k, k \in K$*), can be expressed as*

$$\begin{aligned} g(\delta^1) =& \frac{(c^e + c^l)}{\sqrt{2\pi}} e^{-z^2} \{\bar{\bar{\sigma}}_1 + \bar{\bar{\sigma}}_2 + \cdots + \bar{\bar{\sigma}}_{|K|}\}, \\ =& \frac{(c^e + c^l)}{\sqrt{2\pi}} e^{-z^2} \{\bar{\sigma}_1 + \sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2} + \cdots + \sqrt{\bar{\sigma}_1^2 + \cdots + \bar{\sigma}_{|K|}^2}\}, \end{aligned} \tag{4.22}$$

where $\Phi(z) = \frac{1}{1+\beta}$.

Proof. For sequence $1 \to 2 \to \cdot \to |K|$, $T_{[k]}$ has variance $\bar{\sigma}_1^2 + \cdots + \bar{\sigma}_{[k]}^2$. Apply Lemma 8 to each block $k \in K$. □

Proposition 10 analyzes (4.22) to prescribe the optimal block sequence.

Proposition 10. *Let* $B_{[k]} \sim N(\bar{\mu}_{[k]}, \bar{\sigma}_{[k]}^2)$ *for each* $k \in K$*. The optimal sequence with the optimal planned end-times that minimize the sum of expected earliness and lateness (idleness and overtime associated with the last block, respectively) is the smallest-variance-first-rule.*

Proof. Without loss of generality, sequence B's according to smallest variance first and renumber so that $\sigma_{[k-1]} \leq \sigma_{[k]}, k = 2, \ldots, |K|$. Define $T_{[k]} := B_{[1]} + \cdots + B_{[k]}$ with mean $\bar{\bar{\mu}}_{[k]} = \bar{\mu}_{[1]} + \cdots + \bar{\mu}_{[k]}$ and variance $\bar{\bar{\sigma}}_{[k]}^2 = \bar{\sigma}_{[1]}^2 + \cdots + \bar{\sigma}_{[k]}^2$. Swapping the

first two blocks in the sequence without changing the sequence of other blocks, we obtain

For sequence $1 \to 2 \to \cdots \to N$,

$$\hat{Z} = \frac{(c^e + c^l)}{\sqrt{2\pi}} e^{-z^2} \left[\bar{\sigma}_1 + \sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2} + \cdots + \sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2 + \cdots + \bar{\sigma}_{|K|}^2} \right];$$

and, for sequence $2 \to 1 \to \cdots \to N$,

$$\hat{Z} = \frac{(c^e + c^l)}{\sqrt{2\pi}} e^{-z^2} \left[\bar{\sigma}_2 + \sqrt{\bar{\sigma}_2^2 + \bar{\sigma}_1^2} + \cdots + \sqrt{\bar{\sigma}_2^2 + \bar{\sigma}_1^2 + \cdots + \bar{\sigma}_{|K|}^2} \right].$$

Corresponding terms in the two square brackets are the same, except for the first ones. Thus, it can be seen that the SV rule optimally sequences the first two blocks. So, fix the first block in position. In a similar manner, switching the blocks in the second and third positions shows the SV rule optimally sequences these two blocks as well. By comparing successive pairwise switches, the SV rule can be seen to give the optimal permutation of all blocks. □

Proposition 10 shows that the SV rule gives the optimal sequence of blocks when surgery durations are independent and normally distributed. Based on our preliminary analysis, it does not appear possible to obtain a closed from of $g(\delta)$ for block durations that follow a distribution other than the normal. The next section shows that values $E(T - y)^+$ and $E(y - T)^+$ when T follows either the lognormal or the gamma distribution do not differ by much from the values when T follows the normal distribution with the same parameters, so and that one can apply the results from the normal distribution with little error in these other cases.

4.3 Numerical Study for the Objective Function Value

We have argued that using the representative surgery duration is appropriate for tactical planning purposes. In other contexts in which other surgery durations are appropriate, the lognormal or the gamma distribution may provide a better fit. If block times are right-skewed, perhaps, representing the block times of a series of surgery durations that follow the lognormal distribution. A special case would be a block with a single surgery that follows the lognormal distribution. This section compares expected earliness and lateness values that result if durations follow normal, lognormal, or gamma distributions. This comparison must be done numerically because it appears that a closed form solution can only be obtained for the normal distribution.

Some studies (May et al., 2000; Strum et al., 2000a,b, 2003) have concluded that the lognormal distribution fits actual surgery-duration data well. Depending upon parameter values, the gamma distribution can be right-skewed, similar to the lognormal. We include the gamma distribution in our study to compare both of these right-skewed distributions. Although the normal distribution is analytically tractable, closed form of solutions associated with other distributions (e.g., lognormal, gamma) are not. Thus, we compare the values of lateness $E(T-y)^+$ and earliness $E(y-T)^+$ for each of these distributions (lognormal, gamma and normal). We conduct numerical tests about the values of $E(T-y)^+$ and $E(y-T)^+$ as functions of y for the case of μ=2, 3, 4 and 5, and coefficient of variation (CV) =0.2, 0.3, 0.4 and 0.5. Because all four μ's give similar results, we discuss only the case of $\mu = 4$ and $CV = 0.2$ (i.e., $\sigma = 0.8$), which is displayed in Table 4.2.

Column (1) gives the y values we selected from the range of $(\mu - 3\sigma, \mu + 3\sigma)$; columns (2)-(4) give the expected lateness of lognormal (LN), gamma (G) and normal

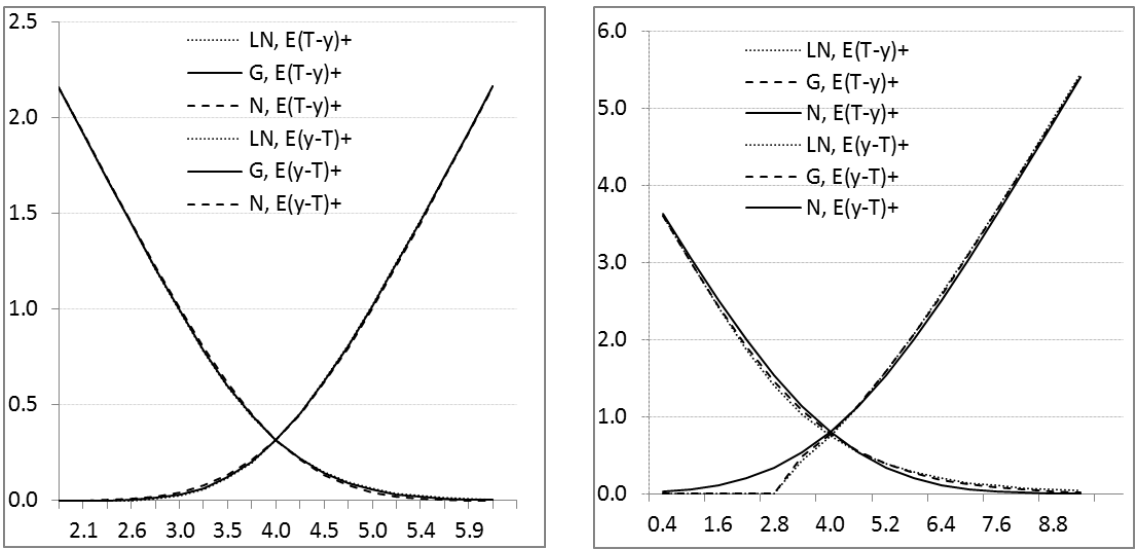
(N) distributions, respectively; column (5) ((6)) gives the difference between the expected lateness of normal and lognormal (gamma) distribution, scaled by μ; columns (7)-(9) give the expected earliness of LN, G and N distributions, respectively; column (10) ((11)) gives the difference between the expected earliness of normal and lognormal (gamma) distribution scaled by μ. Because the relative differences (i.e., columns (5), (6), (10) and (11)) are so small, we assume that three distributions give the same expected earliness and lateness to a close approximation.

Table 4.2: Expected values of Earliness and Lateness When $\mu = 4$ and $\sigma = 0.8$.

	Expected Lateness $E[(T-y)^+]$					Expected Earliness $E[(y-T)^+]$				
y	LN	G	N	$\frac{\lvert(2)-(4)\rvert}{\mu}$	$\frac{\lvert(3)-(4)\rvert}{\mu}$	LN	G	N	$\frac{\lvert(7)-(9)\rvert}{\mu}$	$\frac{\lvert(8)-(9)\rvert}{\mu}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1.84	2.160	2.160	2.161	0.0	0.0	0.000	0.000	0.000	0.0	0.0
2.08	1.920	1.920	1.922	0.1	0.0	0.000	0.000	0.002	0.1	0.0
2.56	1.443	1.445	1.451	0.2	0.2	0.000	0.005	0.011	0.3	0.2
3.28	0.779	0.787	0.800	0.5	0.3	0.059	0.067	0.080	0.5	0.3
3.52	0.596	0.603	0.615	0.5	0.3	0.116	0.123	0.135	0.5	0.3
4.0	0.316	0.318	0.319	0.1	0.0	0.316	0.318	0.319	0.1	0.0
4.48	0.148	0.144	0.135	0.3	0.2	0.628	0.624	0.615	0.3	0.2
5.20	0.039	0.034	0.023	0.4	0.3	1.239	1.234	1.223	0.4	0.3
5.68	0.014	0.011	0.005	0.2	0.1	1.694	1.691	1.685	0.2	0.1
6.16	0.005	0.003	0.001	0.1	0.1	2.165	2.163	2.161	0.1	0.1

Figure 4.5 shows that the graphs of $E(T-y)^+$ and $E(y-T)^+$ are nearly the same for all three of these distributions. Lateness $E(T-y)^+$ is a decreasing function of y, and earliness $E(y-T)^+$ is a increasing function of y. When values of expected earliness and tardiness are not small (e.g., given $2.56 \leq y \leq 5.20$), all three distributions incur approximately the small expected earliness and tardiness as shown by the relatively small differences between distributions. We use μ as a denominator to compare the relative differences, which are parameter-sensitive. When both values are small (e.g., $y = 1.84, 2.08, 5.68$, and 6.16), the absolute differences between expected earliness (and tardiness) associated with the three distributions are small. For example, in case of $y = 2.08$ in Table 4.2, columns (10) and (11) give small values

that are nearly the same, but look different in Figure 4.5 (b).



(a) $\mu = 4, \sigma = 0.8$ (b) $\mu = 4, \sigma = 2$

Figure 4.5: Expected Earliness and Lateness

Because the values of expected lateness and earliness are nearly the same for each of these three distributions, we recommend applying the closed form solution associated with the normal distribution as a close approximation to cases involving either the gamma or lognormal distribution.

4.4 Extensions: No-Shows

Patient no-shows play a major role in deteriorating schedule performance (Lin et al., 2011) because the no-show rate can be significant; for example, they have been reported to be from 22% to more than 50% (Guse et al., 2003) in health-care clinics. Surgery-patient no-shows may result from immediate cancellations before scheduled surgery, due, for example, to failure of patients to prepare for surgery as instructed. Hospital managers can overbook patients to minimize the expected idle time caused by no-shows or employ the following analysis to manage planned block durations

appropriately.

Let α denote the probability of a no-show, a discrete event, and $h(d_i)$ denote the probability distribution function (p.d.f.) for D_i, the duration of a *representative* surgery of sub-specialty i. Define a new p.d.f., $h'(d_i)$ with discrete mass representing a no-show and continuous random duration as follows:

$$h'(d_i) = \begin{cases} \alpha & \text{if } d_i = 0, \\ (1-\alpha)h(d_i) & \text{if } d_i > 0. \end{cases}$$

The associated distribution function of surgery duration, considering the possibility of a no-show, $H'(x)$, is defined as $H'(x) := \alpha + (1-\alpha)H(x)$, where $H(x)$ is the distribution function of $h(x)$. We have to use Lebesgue integration rather than Riemann integration to form $H'(x)$, the distribution function of surgery duration with the possibility of a no-show (Folland, 1999), because Riemann integration for a no-show event is 0. For a single block, we can find the optimal duration $\hat{x}$ as the value at which distribution function $H'(x)$ is equal to the critical ratio:

$$H'(\hat{x}) = \alpha + (1-\alpha)H(\hat{x}) = \frac{1}{1+\beta}.$$

Let $\hat{x}^O$ and $\hat{x}^N$ denote the optimal solutions for the original case without no-shows and the new case with no-shows, respectively. The corresponding distribution functions are given by:

$$H(\hat{x}^O) = \frac{1}{1+\beta}, \tag{4.23}$$

and

$$H(\hat{x}^N) = \frac{1-\alpha-\alpha\beta}{1-\alpha-\alpha\beta+\beta}. \tag{4.24}$$

Figure 4.6 gives the values of $H(\hat{x})$ with ranges of $0 \leq \alpha \leq .3$ and $0.5 \leq \beta \leq 1.5$

to show how optimal block duration changes as a function of α and β. $H(\hat{x})$ is a decreasing function of both α and β. When there are no-shows, we may increase the number of patients scheduled in a given block or decrease the optimal block duration for a given number of patients.

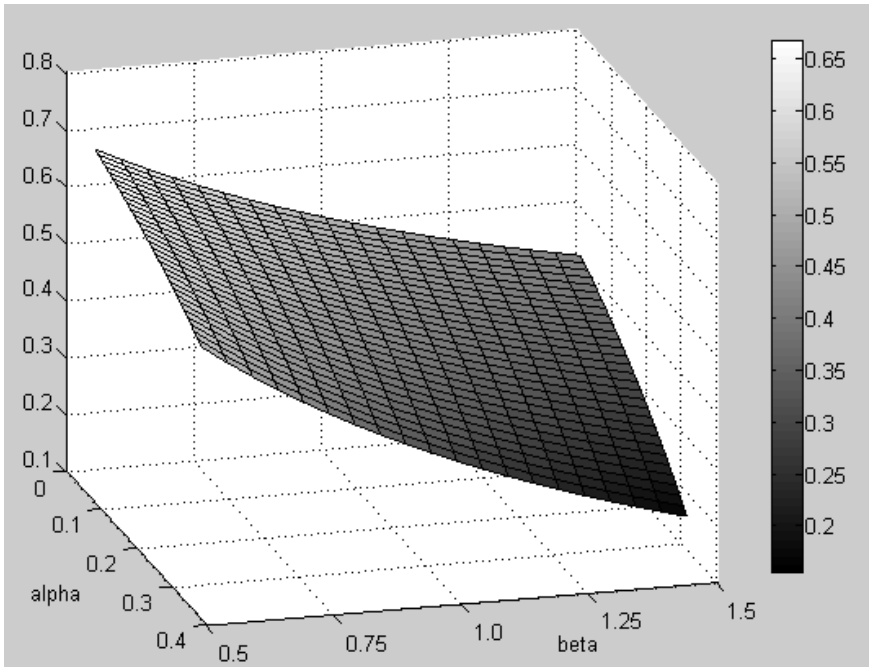


Figure 4.6: Optimal Block Durations with No-show and without No-show

4.5 Managerial Insights

This paper provides managerial insights into MSS, based on the assumptions that forecasts provide the expect number of surgeries to be performed by each surgical subspecialty, that a representative surgery-duration distribution that is normally distributed (according to the CLT) can be derived for each subspecialty based on historical data, that all surgery durations are mutually independent, and that each surgery begins when the previous one ends. Our analysis results in an easy way to compute the optimal planned duration (equivalently, planned end time) of each

time block and shows that time blocks can be optimally sequenced using the easy-to-implement SV rule.

If each subspecialty were responsible for setting the planned duration of its block, the uncertainty in surgery duration might be neglected, resulting in a naive planned block duration equal to the sum of the expected durations of its surgeries. This would parallel the current practice of scheduling the starting time of each surgery to be the sum of the expected durations of surgeries that precede it. Alternatively, each subspecialty might take a myopic approach, neglecting the impact of other subspecialties on the schedule because they do not exchange information, but considering uncertainty by applying a newsvendor model to set planned block duration, say $x'_{[k]}, k \in K$ according to

$$x'_{[k]} = \bar{\mu}_{[k]} + z\bar{\sigma}_{[k]}$$

such that $\Phi(z) = \frac{1}{1+\beta}$.

In contrast, the planned block durations (equivalently planned end times) that our method prescribes deals optimally with uncertainty and depends upon β, the ratio of earliness-to-lateness cost penalties. If $\beta = 1$ (i.e., $c^e = c^l$), the optimal block durations for a given permutation can be specialized to $\hat{x}_{[1]} = \bar{\mu}_{[1]}, \hat{x}_{[2]} = \bar{\mu}_{[2]}, \ldots, \hat{x}_{[|K|]} = \bar{\mu}_{[|K|]}$. This case actually corresponds to the naive approach and shows that it is actually optimal if $\beta = 1$. If $\beta < 1$ (i.e., $c^e < c^l$), z is positive. In other words, if the penalty cost of lateness is greater than that of earliness, the block duration is longer than in the case of $\beta = 1$ (i.e., $\bar{\mu}_{[k]}, k \in K$) to minimize the risk of delaying the next block. In this case, the planned block duration that our method would prescribe, $\hat{x}_{[k]}$, would be less than the duration that the myopic method would prescribe, $x'_{[k]}$ (i.e., $\hat{x}_{[k]} < x'_{[k]}$), indicating that our method is better able to manage the risk of delaying the next block. If $\beta > 1$ (i.e., $c^e > c^l$), z is negative and the

block duration is shorter than in the case of $\beta = 1$ to minimize the risk of idleness. In this case, our method prescribes planned block durations that are longer than the myopic approach (i.e., $\hat{x}_{[k]} > x'_{[k]}$), indicating that our method is better able to deal with the *snowball effect* created by variances accumulating for successive blocks. We now formalize the relationship between the planned block durations that our method and the myopic method prescribe.

Proposition 11. *Consider the planned block duration for kth block as prescribed by our method, $x_{[k]}$, and the myopic method, $x'_{[k]}$.*

$$\hat{x}_{[k]} \leq x'_{[k]} \ \textit{if } \beta \leq 1$$
$$\hat{x}_{[k]} > x'_{[k]} \ \textit{otherwise.}$$

Proof. We use the following fundamental relationship for both cases (i) and (ii):

$$\sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k]}} \leq \sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k-1]}} + \bar{\sigma}_{[k]}.$$

Case (i) $\beta \leq 1$. In this case, $c^e \leq c^l$; i.e., $z \geq 0$.

$$\begin{aligned}
\sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k]}} - \sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k-1]}} \quad &\leq \bar{\sigma}_{[k]} \\
\bar{\mu}_{[k]} + z\big(\sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k]}} - \sqrt{\bar{\sigma}^2_{[1]} + \cdots + \bar{\sigma}^2_{[k-1]}}\big) &\leq \bar{\mu}_{[k]} + z\bar{\sigma}_{[k]} \\
\hat{x}_{[k]} \quad &\leq x'_{[k]}.
\end{aligned}$$

Case (ii) $\beta > 1$. In this case $c^e > c^l$; i.e., $z < 0$. The proof parallels that of case

(i). □

The mean and variance of $T_{[k]}$, which determine the expected earliness and lateness of the last block, would be the same no matter which subspecialty is put in that sequence position; i.e., $f_{[|K|]}(\hat{y}^{\hat{\delta}}_{[|K|]}) = f_{[|K|]}(\hat{y}^{\delta}_{[|K|]}), k \in K, \delta \in \Delta$. Thus, the planned end time of block $[|K|]$, $\hat{y}_{[|K|]}$, which is the planned number of OR hours for the day, does not depend on the sequence. In other words, the subspecialty with largest variance *comes for free* in the last sequence position but would add to total cost if it displaced another subspecialty with a lower variance in an earlier sequence position.

To hedge no shows, a primary question is whether planned block durations should be lengthened or reduced. Our approach is different from an overbooking policy that defines the optimal number of surgeries in a given block time, because we seek the optimal block duration, given the forecast number of surgeries. Considering no-shows reduces optimal block duration in comparison with the case without no-shows. A hospital manager can apply criterion (4.24) to prescribe optimal planned block durations to hedge no-shows.

5. SEQUENCING SURGERIES IN A BLOCK

Surgeon-and-patient-waiting- and OR-idle-times are main sources of inefficiency (Weiss, 1990; Wang, 1993, 1997). Inept sequences that cause excessive amounts of overtime demoralize surgery teams and increase hospital costs. We study the objective of minimizing the sum of the costs of the surgeon-and-patient-waiting- and OR-idle-times analytically, and include any overtime (e.g., amount paid to surgery team members) explicitly incurred in numerical evaluations. Henceforth, we use the terms waiting-, idle-, and over-times, abbreviating these more descriptive phrases to facilitate presentation.

The remainder of this chapter is organized as follows. Section 5.1 gives assumptions and preliminary results, which we apply subsequently. Section 5.2 devises results for cases in which durations follow the lognormal, gamma, or normal distribution to address research objectives (1)-(3). Section 5.3- 5.5 address research objectives (4)-(6) respectively. Section 5.3 analyzes the lognormal in combination with the gamma or with the normal distribution. Section 5.4 extends to the three-surgery case applying numerical results from the two-surgery case. Section 5.5 proposes a heuristic to schedule multiple ORs.

5.1 Preliminaries

This section comprises five subsections. The first three describe our assumptions about patient arrival and ready times, surgery duration, and performance measures, respectively. The next subsection introduces notation and the objective function. The last subsection analyzes some basic relationships.

5.1.1 Patient Arrival and Ready Times

We assume that a patient arrives punctually at the time appointed by the scheduler, following Cayirli and Veral (2003), Kaandorp and Koole (2007) and other studies. Gupta (2007) also assumed that surgeons, other surgery team members, and all patients arrive punctually at specified times, but that $h = 0$ so that surgeries must be scheduled as early as possible in the day (i.e., time block). Both Gupta (2007) and Pinedo (2009) assumed that each patient is ready at the expected completion time of the previous surgery.

Kanich and Byrd (1996) described the scheduling of patient arrival times according to surgery specialty: anesthesia types and genitourinary patients must arrive 1.5 hours and 2 hours before their scheduled starting times, respectively; and others, 1 hour. The OR scheduler determines the scheduled start time for patient j based on the expected completion time of the previous surgery and then directs the patient to arrive at time $t_j - r_s$, where r_s is the time required to complete pre-operative activities for specialty s.

5.1.2 Surgery Duration

In general, studies have assumed that surgery durations are *i.i.d.*; in particular, numerous studies have assumed that surgery durations are exponentially distributed (Cayirli and Veral, 2003) so that models are tractable. A number of studies (May et al., 2000; Strum et al., 2000a,b, 2003) have concluded that the lognormal distribution fits actual surgery-duration data well. After examining a large set of actual surgery-duration data and testing the fit of both lognormal and normal distributions, Strum et al. (2000a) concluded that the lognormal, which is skewed to the right (Casella and Berger, 2001), fits actual data better than the normal. However, not all studies reinforce this conclusion. Tiwari and Berger (2010) found that no

single distribution fits a wide range of surgery duration, and that the lognormal distribution actually fits relatively few actual durations. Stepaniak et al. (2009) investigated the possible dependence of surgery duration on factors like age, surgeon's experience, and team composition.

Depending upon parameter values, the gamma distribution can be right-skewed, similar to the lognormal. We include the gamma distribution in our study to compare both of these right-skewed distributions. Chakraborty et al. (2010) used the gamma distribution to match the mean and variance of the lognormal distribution. The normal distribution is symmetric and has been used commonly in analytical approaches because of its tractability (Casella and Berger, 2001) and general applicability. We compare and contrast the normal and lognormal distributions.

We assume that, once ready, the patient must complete the surgery. We allow the second surgery to start if the first surgery ends after h, because the second patient is ready at the expected completion time of the first one, μ_1. If the second surgery starts after h, it will incur waiting time as well as overtime.

5.1.3 Performance Measures

Some papers have employed only expected waiting- and idle-time penalties; others, only the expected overtime penalty; yet others, all three. Weiss (1990), Wang (1993) and Wang (1997) used the sum of expected waiting- and idle-time penalties. Denton et al. (2010) ignored expected waiting- and idle-time penalties in favor of expected overtime penalty. Gupta (2007), Kaandorp and Koole (2007), Gupta and Denton (2008) and Denton and Gupta (2003) considered all three measures.

If the last surgery in a time block finishes before time h, we ignore this end-of-block idle time because, if it were penalized in the objective function, surgeries could be purposely scheduled later in the block to reduce it, undesirably increasing

the likelihood of incurring overtime. Further, end-of-block idle time could also be reduced by scheduling additional surgeries in the block; however, this would also increase the likelihood of incurring overtime.

We analyze waiting- and idle-time for each of the three distributions and end-of-block overtime for distributions that are tractable, resorting to a numerical tests in cases for which end-of-block overtime cannot be expressed in closed form.

5.1.4 Notation

Let patient j be ready at time t_j for a surgery of random duration X_j with mean μ_j and variance σ_j^2 and consider sequencing patients $j = 1, 2$ in a time block of h hours. Without loss of generality, consider the sequence in which patient 1 precedes patient 2: $X_1 \rightarrow X_2$, where X_1 and X_2 denote the independent and random surgery durations of patients 1 and 2, respectively.

Let $Z_{1,2}^{t_2}$ denote the objective function value for the case in which the sequence of surgeries is 1,2; patient 1 is ready at time $t_1 = 0$; and patient 2, at time t_2. Tardiness $W_{1,2}^2 := (X_1 - t_2)^+$ corresponds to the waiting time associated with the second surgery. Earliness $I_{1,2}^2 := (t_2 - X_1)^+$ corresponds to the idle time associated with the second. Neither waiting- nor idle-time is associated with the first surgery, (i.e., $W_{1,2}^1 = I_{1,2}^1 = 0$) because $t_1 = 0$ and this surgery starts at time 0. Tardiness beyond the end of the block time corresponds to the overtime $O_{1,2} := [\max(X_1, t_2) + X_2 - h]^+$. The subscript on each of these symbols indicates the sequence of surgeries, the superscripts on W and I indicate the surgery associated with the waiting- and idle-time, and superscript on Z indicates the second surgery is scheduled to begin a time t_2.

The analysis in this chapter involves costs per unit time for waiting c^w, idleness c^i, and overtime c^o. Overtime cost is paid explicitly to the surgery team by the hospital, but waiting and idleness costs are accrued implicitly as penalty costs, reflecting

inefficiencies. The objective function for sequence $X_1 \rightarrow X_2$, $Z_{1,2}^{t_2}$, is defined as (5.1):

$$Z_{1,2}^{t_2} = c^w E[W_{1,2}^2] + c^i E[I_{1,2}^2] + c^o E[O_{1,2}]. \tag{5.1}$$

We analyze the sum of expected waiting- and idle-time penalties (SWIP), $c^w E[W_{1,2}^2] + c^i E[I_{1,2}^2]$, analytically and study the expected overtime penalty (OTP), $c^o E[O_{1,2}]$, numerically in subsequent sections.

5.1.5 Analysis of Basic Relationships

We consider an extreme case in which the second patient arrives so early that s/he is ready at time 0, and the surgeon for the second patient is also ready at time 0. For example, a group of patients scheduled for cataract surgery may be directed to arrive at the same time. In this case, which provides a bound, the objective function, $Z_{1,2}^{t_2=0}$ specializes to (5.2):

$$\begin{aligned} Z_{1,2}^{t_2=0} &= c^w E(X_1)^+ + c^i E(0 - X_1)^+ + c^o E[\max(X_1, 0) + X_2 - h]^+ \\ &= c^w E[X_1] + c^o E(X_1 + X_2 - h)^+. \end{aligned} \tag{5.2}$$

The expected overtime, $E(X_1 + X_2 - h)^+$, is independent of the sequence, because it depends only on $X_1 + X_2$. The objective function value, $Z_{1,2}^{t_2=0}$ is increasing in $E(X_1)$, the mean duration of the first surgery. Thus, equation (5.2) shows that the SM rule minimizes SWIP when both ready times are 0.

If one considers scheduling the starting time of the second surgery, t_2, and deal only with SWIP, as Weiss (1990) did, $Z_{1,2}^{t_2}$ must be minimized with respect to (w.r.t.) t_2:

$$Z_{1,2}^{t_2} = c^w \int_{t_2}^{\infty} (X_1 - t_2) f_{X_1}(x_1) dx_1 + c^i \int_{-\infty}^{t_2} (t_2 - X_1) f_{X_1}(x_1) dx_1,$$

This is the objective function of the newsvendor problem, for which the optimal ready time for patient 2 is t_2^* such that $F_{X_1}(t_2^*) = c^w/(c^w + c^i)$ (Weiss, 1990), where $F_{X_1}(t_2^*)$ is the cumulative distribution function of random duration X_1 evaluated at $X_1 = t_2^*$. If $c^w = c^i$ and X_1 is described by the normal distribution, $F_{X_1}(t_2^*) = 0.5$, which means that optimal ready time t_2^* is μ_1, the expected completion time of the first patient, and that there is 50 percent chance of incurring both waiting- and idle-times.

We now introduce a result for the general case in which $t_2 = \mu_1$. Instead of considering t_2^* as a decision variable, $t_2 = \mu_1$ is specified. We invoke this result in subsequent analysis.

Proposition 12. *By definition of partial expected value, expected waiting time- and idle- times associated with the second surgery are equal, i.e.,* $E[W_{1,2}^2] = E[I_{1,2}^2]$.

Proof. See the Appendix. □

Again, with $t_2 = \mu_1$, objective function (5.1) can be simplified as (5.3) by applying Proposition 12.

$$Z_{1,2}^{\mu_1} = (c^w + c^i)E[W_{1,2}^2] + c^o E[O_{1,2}]. \tag{5.3}$$

We do not treat t_2 as a decision variable; rather, we assume that the scheduler uses a simple rule as Gupta (2007) and Pinedo (2009) did, setting $t_2 = \mu_1$, the expected completion time of the first surgery. In numerical tests in Section 5.2, we compare expected overtime with only expected waiting time, since $E[W_{1,2}^2] = E[I_{1,2}^2]$ by Proposition 12.

5.2 Analysis By Probability Distribution

Sequencing two surgeries can provide basic results that lend insights into larger stochastic scheduling problems. Rules applicable to the two-surgery scheduling prob-

lem (Gupta, 2007; Pinedo, 2009) may provide a foundation that can be extended to the general N-surgery case(Weiss, 1990). We note that two-job problems related to single-machine, flow shop, and job shop configurations have been studied similarly to gain insights (Pinedo, 2008, 2009).

In the following subsections, we analyze three surgery-duration distributions (lognormal, gamma, normal) for the two-surgery case as well as the three-surgery case for the normal. We are able to express expected waiting time $E[W_{1,2}^2]$ or $E[W_{2,1}^2]$ to get SWIP in closed form for each distribution, but expected overtime $E[O_{1,2}]$ or $E[O_{2,1}]$ is intractable. We cannot determine the best sequencing rule from expected waiting time for the lognormal and gamma distributions but can for the normal distribution. Hence, we conduct numerical studies to analyze the effect of OTP in comparison with that of SWIP and to specify the optimal sequencing rule for each distribution.

After analyzing actual hospital data for several years, Strum et al. (2000a) and Stepaniak et al. (2009) reported that mean values of surgery durations range from .5 to 6 hours; coefficient of variation (ρ), up to .5 . The numerical study in this chapter deals with an even broader range of parameter values to cover even more general instances. We restrict the sum of mean surgery durations (i.e., $\mu_1 + \mu_2 \leq h$) to preclude excessive overtime and study 2,205 instances, which are formed by combinations of 9 levels of μ_j, each stated in proportion to block duration h for each j (i.e., $\mu_j = 0.1 \times h, 0.2 \times h, \ldots, 0.9 \times h, j = 1, 2$) and 7 levels of ρ, $(0.1, 0.2, \ldots, 0.7)$ for each distribution. The total number of instances can be computed as $2,205 = 45 \times 7 \times 7$, where 45 is the number of μ_1, μ_2 combinations that are feasible with respect to the $\mu_1 + \mu_2 \leq h$ restriction (see Table 5.1 in the next subsection) and the first (second) 7 represents the number of levels of ρ for the first (second) duration. Of the 2,205 instances, $\mu_1 < \mu_2$ (or $\mu_1 > \mu_2$) in 980 instances, $\mu_1 = \mu_2$ in 245 instances, $\sigma_1 < \sigma_2$ (or $\sigma_1 > \sigma_2$) in 1,057 instances, $\sigma_1 = \sigma_2$ in 91 instances, and $\mu_1 + \mu_2 = h$ in

441 instances. We invoke the restriction $\mu_1 + \mu_2 = h$ at several points in our study because this case gives an upper bound on the amount of expected overtime and even for this worst case in which OTP is larger than in other cases for which $\mu_1 + \mu_2 < h$, SWIP contributes more in determining the optimal sequence than OTP does.

In following sections, we compare sequences $X_1 \to X_2$ with $X_2 \to X_1$, evaluating the difference of objective functions $Z_{1,2}^{\mu_1}$ and $Z_{2,1}^{\mu_2}$, defined by $\Delta Z := Z_{1,2}^{\mu_1} - Z_{2,1}^{\mu_2}$:

$$\begin{aligned} \Delta Z &= (c^w + c^i)\{E[W_{1,2}^2] - E[W_{2,1}^2]\} + c^o\{E[O_{1,2}] - E[O_{2,1}]\} \\ &= c^o \Delta E[W]\Big[\frac{c^w + c^i}{c^o} - \gamma\Big], \end{aligned} \tag{5.4}$$

where $\Delta E[W] = E[W_{1,2}^2] - E[W_{2,1}^2]$, $\Delta E[O] = E[O_{1,2}] - E[O_{2,1}]$, and $\gamma = \Delta E[O]/\Delta E[W]$. We use γ to evaluate the impact of OTP in comparison with that of SWIP in numerical studies. As γ goes to zero, decisions that determine $\Delta E[W]$ specify the optimal sequence. However, as γ increases, the cost ratio $(c^w + c^i)/c^o$ may also influence the objective function.

5.2.1 The Lognormal Distribution

The lognormal distribution has been shown to be a good fit for the durations of many actual surgeries (May et al., 2000; Strum et al., 2000a,b, 2003), reflecting non-negativity and right-skewness characteristics (Strum et al., 2000a). Consider the two parameters, λ and δ, of the lognormal distribution, which are actually the mean and the standard deviation of the associated random variable Y, which follows the normal distribution. $X = e^Y$, has the lognormal distribution. The mean μ and variance σ^2 of X can be expressed in terms of the parameters of the distribution of Y:

$$E[X] = \mu = e^{\lambda + \frac{1}{2}\delta^2} \text{ and } V[X] = \sigma^2 = \mu^2(e^{\delta^2} - 1).$$

Strum et al. (2000a) and Stepaniak et al. (2009) used the shifted lognormal distribution to find a better fit than either the lognormal or normal distribution for some surgery durations. If the distribution used to model surgery duration were shifted to the right by amount s, its shifted mean would be $E(X+s) = \mu + s$ and its variance would be $V(X+s) = V(X)$. Such a location parameter does not influence our analysis, because $E(X+s-\mu-s)^+ = E(X-\mu)^+$ and $E(\mu+s-X-s)^+ = E(\mu-X)^+$.

Consider two surgeries ($j = 1, 2$), each with lognormally distributed duration X_j, mean μ_j, standard deviation σ_j, and associated parameters λ_j and δ_j, where $j = 1, 2$. Proposition 13 establishes the objective function $Z_{1,2}^{\mu_1}$ for a sequence of two such surgeries:

Proposition 13. *The objective function for a sequence of two lognormally distributed durations is given by:*

$$\begin{aligned} Z_{1,2}^{\mu_1} &= (c^w + c^i)E(X_1)\Big[2\Phi(\frac{\delta_1}{2}) - 1\Big] + c^o E[O_{1,2}], \qquad (5.5) \\ \textit{where } E(X_1) &= e^{\lambda_1 + \frac{1}{2}\delta_1^2}. \end{aligned}$$

Proof. See the Appendix. □

Equation (5.5) does not clearly identify a sequencing rule. Both $E(X)\big[2\Phi(\frac{\delta}{2})-1\big]$ and the standard deviation $\sigma = \sqrt{V(X)} = E(X)\sqrt{e^{\delta^2}-1}$ are product forms of $E[X]$ and an increasing function of δ. Hence, we conjecture that the SV rule minimizes SWIP.

We conduct numerical tests to assess which rule, SV or SM, gives better results relative to SWIP in each of the 2,205 instances, leaving OTP for later analysis. We analyze numerical tests in two-dimensional tabular form as follows. Table 5.1 shows that the SV rule gives better results than the SM rule in a meaningful pattern of test

Table 5.1: Comparison of SM and SV for Sequences of Two Lognormally Distributed Surgeries

$\mu_1 \backslash \mu_2$	1	2	3	4	5	6	7	8	9
1	NA, $\frac{42}{42}$	$\frac{40}{49}, \frac{46}{46}$	$\frac{44}{49}, \frac{47}{47}$	$\frac{46}{49}, \frac{48}{48}$	$\frac{47}{49}, \frac{48}{48}$	$\frac{48}{49}, \frac{48}{48}$	$\frac{49}{49}, \frac{48}{48}$	$\frac{49}{49}, \frac{49}{49}$	$\frac{49}{49}, \frac{49}{49}$
2	$\frac{40}{49}, \frac{46}{46}$	NA, $\frac{42}{42}$	$\frac{35}{49}, \frac{47}{47}$	$\frac{40}{49}, \frac{46}{46}$	$\frac{42}{49}, \frac{48}{48}$	$\frac{44}{49}, \frac{47}{47}$	$\frac{45}{49}, \frac{48}{48}$	$\frac{46}{49}, \frac{48}{48}$	-
3	$\frac{44}{49}, \frac{47}{47}$	$\frac{35}{49}, \frac{47}{47}$	NA, $\frac{42}{42}$	$\frac{33}{49}, \frac{47}{48}$	$\frac{37}{49}, \frac{47}{48}$	$\frac{40}{49}, \frac{47}{47}$	$\frac{41}{49}, \frac{48}{48}$	-	-
4	$\frac{46}{49}, \frac{48}{48}$	$\frac{40}{49}, \frac{46}{46}$	$\frac{33}{49}, \frac{47}{48}$	NA, $\frac{42}{42}$	$\frac{31}{49}, \frac{48}{48}$	$\frac{35}{49}, \frac{47}{47}$	-	-	-
5	$\frac{47}{49}, \frac{48}{48}$	$\frac{42}{49}, \frac{48}{48}$	$\frac{37}{49}, \frac{47}{48}$	$\frac{31}{49}, \frac{48}{48}$	NA, $\frac{42}{42}$	-	-	-	-
6	$\frac{48}{49}, \frac{48}{48}$	$\frac{44}{49}, \frac{47}{47}$	$\frac{40}{49}, \frac{46}{46}$	$\frac{35}{49}, \frac{48}{48}$	-	-	-	-	-
7	$\frac{49}{49}, \frac{48}{48}$	$\frac{45}{49}, \frac{48}{48}$	$\frac{41}{49}, \frac{48}{48}$	-	-	-	-	-	-
8	$\frac{49}{49}, \frac{49}{49}$	$\frac{46}{49}, \frac{48}{48}$	-	-	-	-	-	-	-
9	$\frac{49}{49}, \frac{49}{49}$	-	-	-	-	-	-	-	-

instances. The left-most column in Table 5.1 lists values of μ_1; the top-most row, μ_2. Each of 45 cells represents a combination that is feasible w.r.t. the $\mu_1 + \mu_2 \leq h$ restriction. Each cell represents 49 (i.e., 7×7) different variance levels. The left (right) denominator in each cell represents the numbers of the valid instances related to the SM (SV) rule. For example, when $\mu_1 = \mu_2$, the SM rule is not applicable (NA) because both means have the same value; similarly, the SV cannot be applied in 7 cases in which both variances are the same. The left (right) numerator in each cell represents the number of valid instances in which the SM (SV) rule gives the better value of SWIP (each of the two rules gives a SWIP measure - the same values for many instances - and we identify the number on instances out of valid instances which each rule gives the better result, without claiming that it is the optimal result. For example, when $\mu_1 = 4$ and $\mu_2 = 3$ (or vice versa), the SV rule gives the better result for 47 of 48 instances but the SM rule gives the better results for only 33 of 49 instances. In summary, the SV rule gives the better results in instances represented by all cells in Table 5.1; the SM rule ties, giving the better result for many instances represented by most cells, but never gives results that improve on those achieved by

the SV rule.

We now describe our numerical tests, which we designed to evaluate the relative impact of SWIP in comparison with that of OTP. Table 5.2 shows results for a selected sample of 10 instances out of the 2,205 tested. Column (1) gives test instance number, columns (2) and (3) give parameter values of each of two surgeries (i.e., μ_j and σ_j, respectively), columns (4) and (6) give $E[W_{1,2}^2]$ and $E[W_{2,1}^2]$, respectively, columns (5) and (7) give $E[O_{1,2}]$ and $E[O_{2,1}]$, respectively, columns (8)-(10) give measures described in section : $\Delta E[W]$, $\Delta E[O]$, and γ, respectively. The selected sample shows typical cases: in instances 1-8, $\mu_1 < \mu_2$; in instances 1, 3, 5 and 9, $\sigma_1 < \sigma_2$; in instances 2, 6, 7 and 10, $\sigma_1 > \sigma_2$; in instances 4 and 8, $\sigma_1 = \sigma_2$; and in instances 9 and 10, $\mu_1 + \mu_2 = h$.

Both an ANOVA and a t-test show that SV and SM are both statistically significantly effective in minimizing SWIP, each with a $p - value$ of less than 0.0001. We employ simple statistics (e.g., numbers of instances) to compare these two rules in subsequent sections. If $\mu_1 < \mu_2$, sequence $X_1 \to X_2$ is better w.r.t. SWIP than $X_2 \to X_1$ in 841 of 980 instances. If $\sigma_1 < \sigma_2$, sequence $X_1 \to X_2$ is better w.r.t. SWIP than $X_2 \to X_1$ in 1,055 of 1,057 instances. In most instances $\Delta E[O]$ is very small compared to $\Delta E[W]$ (e.g., instances 1-6) so that γ is small and $\Delta E[W]$ determines the best sequence (see (5.4)). In instances for which $\Delta E[O] \simeq \Delta E[W]$, $\Delta E[O]$ is so small that γ is large; for example, instance 8 in Table 5.2. If $|\mu_1 + \mu_2 - h| < \epsilon$ (e.g., instances 7-10), even in this worst case, OTP does not play a dominant role in determining the optimal sequence. Although expected overtime is greater than expected waiting time, $\Delta E[O]$ is less than $\Delta E[W]$. In other words, SWIP (i.e., $\Delta E[W]$) dominates OTP (i.e., $\Delta E[O]$) in all instances for which $(c^w + c^i)/c^o$ is not small (i.e., c^o is not bigger than the sum of other two costs).

Table 5.2: Comparison of $\Delta E[O]$ with $\Delta E[W]$ for Sequences of Two Lognormally Distributed Surgeries

Instance Index (1)	$X_1 \sim$ (μ_1, σ_1) (2)	$X_2 \sim$ (μ_2, σ_2) (3)	$X_1 \to X_2$		$X_2 \to X_1$		Difference		
			$E[W^2_{1,2}]$ (4)	$E[O_{1,2}]$ (5)	$E[W^2_{2,1}]$ (6)	$E[O_{2,1}]$ (7)	$\Delta E[W]$ (8)	$\Delta E[O]$ (9)	$\gamma(\%)$ (10)
1	(1, 0.1)	(2, 0.2)	0.040	0.000	0.080	0.000	-0.040	0.000	0.0
2	(1, 0.6)	(2, 0.4)	0.218	0.000	0.158	0.000	0.061	0.000	0.0
3	(2, 0.2)	(3, 0.3)	0.080	0.000	0.119	0.000	-0.040	0.000	0.0
4	(2, 0.6)	(3, 0.6)	0.233	0.000	0.237	0.000	-0.003	0.000	0.0
5	(3, 0.3)	(4, 0.4)	0.119	0.000	0.159	0.000	-0.040	0.000	0.0
6	(3, 0.9)	(4, 0.8)	0.350	0.011	0.316	0.012	0.035	-0.001	2.3
7	(4, 0.8)	(5, 0.5)	0.316	0.087	0.199	0.098	0.117	-0.011	9.3
8	(4, 2.0)	(5, 2.0)	0.747	0.847	0.764	0.841	-0.017	0.006	37.3
9	(5, 0.5)	(5, 1.0)	0.199	0.513	0.394	0.522	-0.195	-0.009	4.7
10	(5, 3.0)	(5, 2.5)	1.092	1.768	0.934	1.741	0.159	0.027	17.2

For example, $\Delta Z = -0.017(c^w+c^i)+0.006c^o$ for instance 8. If $(c^w+c^i)/c^o > 0.373$, SWIP dominates OTP. Otherwise (i.e., c^o is greater than 2.68 (=1/0.373) times of $c^w + c^i$), SWIP contributes less in determining the optimal sequence than OTP does. Hence, we recommend that the SV rule be used in the two-surgery case in which both durations are lognormally distributed, because it gives better results in the majority of instances, even though it is not globally optimal.

5.2.2 The Gamma Distribution

With certain parameter values, the gamma distribution has a shape similar to the right-skewed form of the lognormal distribution. Because each surgery comprises several small tasks, such as administering anesthesia, performing surgery and closing the wound, the gamma distribution may be used in phase-type distributions to fit service times in such a serial process. The gamma distribution with parameters n and β has mean $E(X) = n\beta$ and variance $V(X) = n\beta^2$. If n is restricted to be an integer, the gamma specializes to the Erlang distribution for which the objective function can be expressed as follows.

Proposition 14. *The total-cost objective function for a sequence of two surgeries, each of which follows a gamma-distributed duration with parameters β_j and integer n_j for $j = 1, 2$, is given by:*

$$Z_{1,2}^{\mu_1} = (c^w + c^i)E(X_1)\frac{n_1^{n_1}}{n_1!}e^{-n_1} + c^o E[O_{1,2}]. \tag{5.6}$$

Proof. See the Appendix. □

Objective function (5.6) does not clearly identify a sequencing rule. Thus, we conduct numerical tests to assess whether the SV rule is better w.r.t. SWIP than the SM rule as it is for the lognormal distribution. Further, using (5.4) we evaluate the

relative impact of SWIP in comparison with that of OTP for the gamma distribution with general n (i.e., not restricted to an integer). Table 5.3 shows results of a selected subset of 10 out of 2,205 instances tested. Each column in Table 5.3 records the same information as the corresponding column of Table 5.2.

If $\mu_1 < \mu_2$, sequence $X_1 \to X_2$ is better w.r.t. SWIP than $X_2 \to X_1$ in 839 of 980 instances. If $\sigma_1 < \sigma_2$, sequence $X_1 \to X_2$ is better w.r.t. SWIP than $X_2 \to X_1$ in 1,056 of 1,057 instances. In most instances $\Delta E[O]$ is very small compared to $\Delta E[W]$ (e.g., instances 1-6) so that γ is small and $\Delta E[W]$ determines the best sequence (see (5.4)). In instances for which $\Delta E[O] \simeq \Delta E[W]$, $\Delta E[O]$ is so small that γ is large (e.g., instance 8 in Table 5.2). If $|\mu_1 + \mu_2 - h| < \epsilon$ (e.g., instances 7-10), even in this worst case, OTP does not play a dominant role in determining the optimal sequence. Although expected overtime is greater than expected waiting time, $\Delta E[O]$ is less than $\Delta E[W]$. In other words, SWIP dominates OTP in all instances for which $(c^w + c^i)/c^o$ is not small. Hence, we recommend that the SV rule be used in the two-surgery case in which both surgery durations are gamma distributed, because it gives better results in the majority of instances, even though it is not globally optimal.

Table 5.3: Comparison of $\Delta E[O]$ with $\Delta E[W]$ for Sequences of Two Gamma Distributed Surgeries

Instance Index (1)	$X_1 \sim$ (μ_1, σ_1) (2)	$X_2 \sim$ (μ_2, σ_2) (3)	$X_1 \to X_2$		$X_2 \to X_1$		Difference		
			$E[W^2_{1,2}]$ (4)	$E[O_{1,2}]$ (5)	$E[W^2_{2,1}]$ (6)	$E[O_{2,1}]$ (7)	$\Delta E[W]$ (8)	$\Delta E[O]$ (9)	$\gamma(\%)$ (10)
1	(1, 0.1)	(2, 0.2)	0.040	0.000	0.080	0.000	-0.040	0.000	0.0
2	(1, 0.6)	(2, 0.4)	0.232	0.000	0.159	0.000	0.073	0.000	0.0
3	(2, 0.2)	(3, 0.3)	0.080	0.000	0.120	0.000	-0.040	0.000	0.0
4	(2, 0.6)	(3, 0.6)	0.238	0.000	0.239	0.000	-0.001	0.000	0.0
5	(3, 0.3)	(4, 0.4)	0.120	0.000	0.159	0.000	-0.040	0.000	0.0
6	(3, 0.9)	(4, 0.8)	0.356	0.007	0.318	0.007	0.038	-0.000	1.0
7	(4, 0.8)	(5, 0.5)	0.318	0.083	0.199	0.093	0.119	-0.011	9.1
8	(4, 2.0)	(5, 2.0)	0.781	0.860	0.787	0.856	-0.006	0.004	75.2
9	(5, 0.5)	(5, 1.0)	0.199	0.518	0.398	0.524	-0.198	-0.006	3.2
10	(5, 3.0)	(5, 2.5)	1.162	1.850	0.977	1.824	0.185	0.030	14.1

5.2.3 The Normal Distribution

The normal distribution is used in many applications because it is relatively mathematically tractable and, due to the Central Limit Theorem (Casella and Berger, 2001), finds wide application. The normal distribution admits negative values, but surgery duration is strictly positive. However, with a coefficient of variation $\sigma_j/\mu_j < 0.2$ for $j = 1, 2$, the probability that a duration would have a negative value is negligible. If surgery duration is determined as the sum of a number of random task times (e.g., breathing tube insertion (i.e., intubation), anesthesia administration, a series of procedures with different current procedure terminology (CPT) codes such as discectomy and foramenotomy in spine surgery, wound closing, OR cleaning) that are independent because, for example, they are performed by different personnel - as in the case of a number of surgery types - its coefficient of variation would most likely satisfy this condition.

Consider two surgeries with normally distributed durations, $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$. The total cost objective function, $Z_{1,2}^{\mu_1}$, cannot be expressed in closed form because the expected overtime term is intractable, but it can be computed numerically:

$$Z_{1,2}^{\mu_1} = (c^w + c^i)E[W_{1,2}^2] + c^o E[O_{1,2}]. \tag{5.7}$$

$E[W_{1,2}^2]$ can be expressed in closed form:

Proposition 15. *For a sequence $X_1 \to X_2$ of two surgeries with normally distributed durations, $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$*

$$E[W_{1,2}^2] = E[I_{1,2}^2] = \frac{\sigma_1}{\sqrt{2\pi}}. \tag{5.8}$$

Proof. See the Appendix. □

In general, $E[O_{1,2}]$ cannot be expressed in closed form, but it can be if either one of two approximations is appropriate: (i) $X_1 + X_2 < h$ a.s., or (ii) $|X_1 + X_2 - h| < \epsilon$. Case (i) incurs no overtime a.s. Case (ii) occurs when the sum of the two surgery durations is close to the end-of-block time h a.s. Proposition 16 establishes that $Z_{1,2}^{\mu_1}$ can be approximated in closed form in each of these cases.

Proposition 16. *The objective function for sequence $X_1 \to X_2$ of two surgeries with normally distributed durations is approximated by :*

$$Z_{1,2}^{\mu_1} = \begin{cases} (c^w + c^i)\frac{\sigma_1}{\sqrt{2\pi}} & \text{if } X_1 + X_2 < h \text{ a.s.} \\ (c^w + c^i)\frac{\sigma_1}{\sqrt{2\pi}} + c^o(\frac{\sigma_1+\sigma_2}{2\sqrt{2\pi}} + \frac{\sigma_1\sigma_2}{2\pi}) & \text{if } |X_1 + X_2 - h| < \epsilon. \end{cases} \tag{5.9}$$

Proof. See the Appendix. □

$E[W_{1,2}^2]$ in (5.9) is an increasing function of σ_1 and not a function of mean μ_1, so the SV rule minimizes SWIP in this case. Numerical tests also show that if $\sigma_1 < \sigma_2$, sequence $X_1 \to X_2$ minimizes w.r.t. SWIP instances of 1,070 instances. We conduct numerical tests to assess the relative impact of OTP on the objective function value when the sum of surgery durations does not satisfy either (i) or (ii). Each column in Table 5.4 records the same information reported by the corresponding column of Table 5.2. Table 5.4 shows numerically that the SV rule gives better SWIP in all cases. Instances 1 - 8 represent case (i), for which no overtime is incurred. Instances 9 and 10 represent case (ii), for which OTP contributes less in determining the optimal sequence than SWIP does. In these cases, although expected overtime is greater than expected waiting time, $\Delta E[O]$ is less than $\Delta E[W]$. If $\sigma_1 > \sigma_2$, sequence $X_1 \to X_2$ (i.e., largest-variance-first-rule (LV)) is better w.r.t. OTP than $X_2 \to X_1$ in 623 of 1,070 instances (e.g., instance 7); and the values of OTP have no difference in the remaining 439 of 1,070 instances.

Table 5.4: Comparison of $\Delta E[O]$ with $\Delta E[W]$ for Sequences of Two Normally Distributed Surgeries

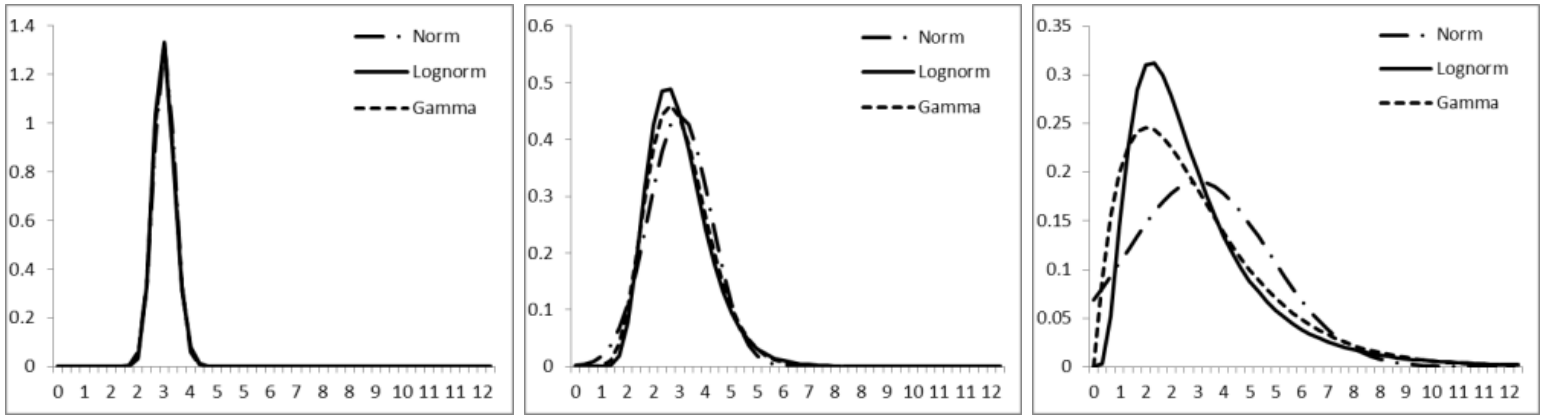
Instance Index (1)	$X_1 \sim$ (μ_1, σ_1) (2)	$X_2 \sim$ (μ_2, σ_2) (3)	$X_1 \to X_2$		$X_2 \to X_1$		Difference		
			$E[W^2_{1,2}]$ (4)	$E[O_{1,2}]$ (5)	$E[W^2_{2,1}]$ (6)	$E[O_{2,1}]$ (7)	$\Delta E[W]$ (8)	$\Delta E[O]$ (9)	$\gamma(\%)$ (10)
1	(1, 0.1)	(2, 0.2)	0.040	0.000	0.080	0.000	-0.040	0.000	0.0
2	(1, 0.6)	(2, 0.4)	0.239	0.000	0.160	0.000	0.080	0.000	0.0
3	(2, 0.2)	(3, 0.3)	0.080	0.000	0.120	0.000	-0.040	0.000	0.0
4	(2, 0.6)	(3, 0.6)	0.239	0.000	0.239	0.000	0.000	0.000	0.0
5	(3, 0.3)	(4, 0.4)	0.120	0.000	0.160	0.000	-0.040	0.000	0.0
6	(3, 0.9)	(4, 0.8)	0.359	0.003	0.319	0.003	0.040	0.000	0.0
7	(4, 0.8)	(5, 0.5)	0.319	0.072	0.199	0.082	0.120	-0.010	8.2
8	(4, 2.0)	(5, 2.0)	0.798	0.824	0.798	0.824	0.000	0.000	0.0
9	(5, 0.5)	(5, 1.0)	0.199	0.522	0.399	0.522	-0.199	0.000	0.0
10	(5, 3.0)	(5, 2.5)	1.197	1.876	0.997	1.876	0.200	0.000	0.0

For two normally distributed surgery durations, each of which is symmetric and bell-shaped, Proposition 16 establishes analytically that both two sequences $X_1 \to X_2$ and $X_2 \to X_1$ give the same expected overtime when $h = \mu_1 + \mu_2$ (441 instances). A numerical study for two normally distributed surgeries shows that 421 of 441 instances have no difference in expected overtime and the remaining 20 instances have little difference.

5.3 Lognormal in Combination with Another Distribution

We assume that two surgeries follow the same distribution in previous sections; however, two surgeries may follow different distributions, for example, because the ages of the patients and/or the experience levels of surgeons are different. In this section, we consider the lognormal in combination with other distributions.

Figure 5.1 shows probability distribution function of each of the three distributions with a common $\mu = 3$ but three different levels of ρ, as a typical example. When ρ is small as in Figure 5.1 (a), all three distributions have the same shape and their graphs look as one because probability functions differ little. As ρ increases in Figures 5.1 (b) and 5.1 (c), the lognormal and gamma distributions become more right-skewed and continue to look like each other but less like the normal. However, we expect that most surgery duration distributions have coefficients of variations at the smaller end of this range of ρ values.



(a) $\mu = 3, \rho = 0.1$ (b) $\mu = 3, \rho = 0.3$ (c) $\mu = 3, \rho = 0.7$

Figure 5.1: Comparison of the Shapes of Distributions with Common Mean = 3.

We compare $E[(X-\mu)^+]$ values for lognormal, gamma and normal distributions because this term has a significant impact on determining the optimal sequence. Our numerical tests involve 9 levels of μ and 7 levels of ρ as Section 5.2. Table 5.5 compares values of $E[W_{1,2}^2]$ as a function of μ and σ to evaluate SWIP for each of these distributions. Column (1) gives instance index; column (2) gives parameters (μ, σ) tested; column (3) gives ρ; columns (4), (5), and (6) give $E[W_{1,2}^2]$ for lognormal, gamma, and normal distributions, respectively; and three right most columns give the relative difference of $E[W_{1,2}^2]$ values for each pair of distributions. Numerical tests show that these relative differences depend on the value of ρ and are increasing functions of ρ. Our analysis strongly suggests that lognormal, gamma, and normal distributions all give similar values of $E[W_{1,2}^2]$, leading us to conjecture that the SV rule is effective relative to SWIP when the lognormal is combined with either the gamma or the normal and, more generally, to the conjecture that any particular distribution analyzed gives results that are similar for all three so that the most convenient (i.e., tractable) distribution can be used in typical cases.

In the next subsections, we study combinations of the lognormal with either the gamma or the normal distribution. Numerical tests are designed to assess the efficacy of the SV rule relative to SWIP and to evaluate the relative impact of OTP in comparison with that of SWIP.

Table 5.5: Comparison of Expected Waiting Times by Surgery Duration

Instance Index (1)	Parameter values (μ, σ) (2)	ρ (3)	$E[W_{1,2}^2]$ Lognormal (4)	Gamma (5)	Normal (6)	Relative Difference $\frac{(5)-(4)}{(4)}(\%)$	$\frac{(6)-(4)}{(4)}(\%)$	$\frac{(6)-(5)}{(5)}(\%)$
1	(1, 0.1)	0.1	0.040	0.040	0.040	0.2	0.3	0.1
2	(1, 0.3)	0.3	0.117	0.119	0.120	1.8	2.6	0.7
3	(1, 0.5)	0.5	0.187	0.195	0.199	4.6	6.8	2.1
4	(1, 0.7)	0.7	0.248	0.268	0.279	8.2	12.7	4.1
5	(2, 0.2)	0.1	0.080	0.080	0.080	0.2	0.3	0.1
6	(2, 0.6)	0.3	0.233	0.238	0.239	1.8	2.6	0.7
7	(2, 1.0)	0.5	0.373	0.391	0.399	4.6	6.8	2.1
8	(2, 1.4)	0.7	0.496	0.536	0.559	8.2	12.7	4.1
9	(5, 0.5)	0.1	0.199	0.199	0.199	0.2	0.3	0.1
10	(5, 1.5)	0.3	0.583	0.594	0.598	1.8	2.6	0.7
11	(5, 2.5)	0.5	0.934	0.977	0.997	4.6	6.8	2.1
12	(5, 3.5)	0.7	1.239	1.341	1.396	8.2	12.7	4.1

5.3.1 Lognormal in Combination with the Gamma Distribution

We consider one surgery with duration that follows the lognormal distribution in combination with another that follows the gamma distribution, noting that both distributions may have a similar shape for selected parameter values. The number of instances and parameter values are the same as ones used in Section 5.2.

Table 5.6 gives results of our numerical tests, which show that the SV rule is better w.r.t. SWIP than the SM rule, and that OTP contributes less in determining the optimal sequence than SWIP does. When the variance of the lognormal is less than the variance of the gamma, scheduling the lognormal duration first (i.e., according to SV) is better than the alternative sequence w.r.t. OTP in all 1,057 instances. When the variance of the gamma is less than that of the lognormal, scheduling the gamma duration first (i.e., according to SV) is better than the alternative sequence w.r.t. SWIP in 1,055 out of 1,057 instances, and SWIP contributes more in determining the optimal sequence than OTP does in 2,184 of 2,205 instances. In the remaining 21 instances, $\Delta E[W] \simeq \Delta E[O]$. If $|\mu_1 + \mu_2 - h| < \epsilon$ (e.g., instances 7-10), even in this worst case, OTP does not play a dominant role in determining the optimal sequence. Although expected overtime is greater than expected waiting time, $\Delta E[O]$ is less than $\Delta E[W]$. In other words, SWIP dominates OTP in all instances for which $(c^w + c^i)/c^o$ is not small. In instances for which $\Delta E[O] \simeq \Delta E[W]$, $\Delta E[O]$ is so small that γ is large (e.g., instance 10 in Table 5.6, see (5.4)).

Table 5.6: Comparison of $\Delta E[O]$ and $\Delta E[W]$ for Sequences of Lognormally(LN) and Gamma(G) Distributed Surgeries

Instance Index (1)	$X_1 \sim$ LN(μ_1, σ_1) (2)	$X_2 \sim$ G(μ_2, σ_2) (3)	$X_1 \to X_2$ $E[W_{1,2}^2]$ (4)	$E[O_{1,2}]$ (5)	$X_2 \to X_1$ $E[W_{2,1}^2]$ (6)	$E[O_{2,1}]$ (7)	Difference $\Delta E[W]$ (8)	$\Delta E[O]$ (9)	$\gamma(\%)$ (10)
1	(1, 0.1)	(2, 0.2)	0.040	0.000	0.080	0.000	-0.040	0.000	0.0
2	(1, 0.6)	(2, 0.4)	0.218	0.000	0.159	0.000	0.060	0.000	0.0
3	(2, 0.2)	(3, 0.3)	0.080	0.000	0.120	0.000	-0.040	0.000	0.0
4	(2, 0.6)	(3, 0.6)	0.233	0.000	0.239	0.000	-0.005	0.000	0.0
5	(3, 0.3)	(4, 0.4)	0.119	0.000	0.159	0.000	-0.040	0.000	0.0
6	(3, 0.9)	(4, 0.8)	0.350	0.011	0.318	0.008	0.030	0.003	7.8
7	(4, 0.8)	(5, 0.5)	0.316	0.098	0.199	0.084	0.120	0.014	12.2
8	(4, 2.0)	(5, 2.0)	0.747	0.839	0.787	0.851	-0.040	-0.012	29.2
9	(5, 0.5)	(5, 1.0)	0.199	0.523	0.398	0.513	-0.199	0.011	5.3
10	(5, 3.0)	(5, 2.5)	1.092	1.757	0.977	1.810	0.120	-0.052	45.2

5.3.2 Lognormal in Combination with the Normal Distribution

We now consider a combination of surgery-duration distributions, one lognormal and the other normal. Even though the lognormal is right-skewed and the normal is symmetric, expected waiting times associated with both are nearly the same as shown in Table 5.6. Hence, we conjecture that the SV rule is better than the SM rule w.r.t. SWIP in this case as well.

Table 5.7 shows that the SV rule is better than the SM rule w.r.t. SWIP, and that OTP contributes less in determining the optimal sequence than SWIP does if $(c^w + c^i)/c^o$ is not small. When the variance of the lognormal is less than that of the normal, scheduling the lognormal first (i.e., according to the SV) is better than the alternative sequence w.r.t. SWIP in all 1,057 instances. When the variance of the normal is less than that of the lognormal, scheduling the normal first (i.e., according to the SV) is better than the alternative sequence w.r.t. SWIP in 1,055 of 1,057 instances. SWIP contributes more in determining the optimal sequence than OTP does in 2,177 of 2,205 instances. In the remaining 28 instances, $\Delta E[W] \simeq \Delta E[O]$. If $|\mu_1 + \mu_2 - h| < \epsilon$ (e.g., instances 7-10), even in this worst case, OTP does not play a dominant role in determining the optimal sequence. Although expected overtime is greater than expected waiting time, $\Delta E[O]$ is less than $\Delta E[W]$. In other words, SWIP dominates OTP in all instances for which $(c^w + c^i)/c^o$ is not small. When one surgery duration is lognormally distributed and the other is normally distributed, we recommend that the SV be used because of its efficacy relative to SWIP and OTP, although the SV rule is not optimal globally.

Table 5.7: Comparison of $\Delta E[O]$ with $\Delta E[W]$ for Sequences of Lognormally(LN) and Normally(N) Distributed Surgeries

Instance Index (1)	$X_1 \sim$ LN(μ_1, σ_1) (2)	$X_2 \sim$ N(μ_2, σ_2) (3)	$X_1 \to X_2$ $E[W^2_{1,2}]$ (4)	$E[O_{1,2}]$ (5)	$X_2 \to X_1$ $E[W^2_{2,1}]$ (6)	$E[O_{2,1}]$ (7)	Difference $\Delta E[W]$ (8)	$\Delta E[O]$ (9)	$\gamma(\%)$ (10)
1	(1, 0.1)	(2, 0.2)	0.040	0.000	0.080	0.000	-0.040	0.000	0.0
2	(1, 0.6)	(2, 0.4)	0.218	0.000	0.160	0.000	0.059	0.000	0.0
3	(2, 0.2)	(3, 0.3)	0.080	0.000	0.120	0.000	-0.040	0.000	0.0
4	(2, 0.6)	(3, 0.6)	0.233	0.000	0.239	0.000	-0.006	0.000	0.0
5	(3, 0.3)	(4, 0.4)	0.119	0.000	0.160	0.000	-0.040	0.000	0.0
6	(3, 0.9)	(4, 0.8)	0.350	0.009	0.319	0.005	0.031	0.004	14.2
7	(4, 0.8)	(5, 0.5)	0.316	0.097	0.199	0.075	0.116	0.022	18.9
8	(4, 2.0)	(5, 2.0)	0.747	0.809	0.798	0.818	-0.051	-0.009	17.4
9	(5, 0.5)	(5, 1.0)	0.199	0.520	0.399	0.510	-0.200	0.010	5.1
10	(5, 3.0)	(5, 2.5)	1.092	1.731	0.997	1.788	0.095	-0.057	60.3

5.4 Three Surgeries

In describing procedures at a local hospital, our health care collaborator emphasized that, typically, only two or three surgeries are scheduled in each OR each day. We provide analytical expression for the three-surgery case. For three surgeries, let $Z_{1,2,3}^{t_2,t_3}$ denote the objective function value for sequence $X_1 \to X_2 \to X_3$ with successive patient ready times $t_1 = 0, t_2$ and t_3. Waiting times $W_{1,2,3}^2 := (X_1 - \mu_1)^+$ and $W_{1,2,3}^3 := [\max(X_1, \mu_1) + X_2 - \mu_1 - \mu_2]^+$ correspond to the second and third surgeries, respectively. Idle times $I_{1,2,3}^2 := (\mu_1 - X_1)^+$ and $I_{1,2,3}^3 := [\mu_1 + \mu_2 - \max(X_1, \mu_1) - X_2]^+$ correspond to the second and third surgeries, respectively. Neither waiting- nor idle-time is associated with the first surgery, (i.e., $W_{1,2,3}^1 = I_{1,2,3}^1 = 0$) because $t_1 = 0$ and this surgery starts at time 0. $O_{1,2,3} := \{\max[\max(X_1, \mu_1) + X_2, \mu_1 + \mu_2] + X_3 - h\}^+$.

Consider three random surgery durations, X_1, X_2, and X_3. $E[W_{1,2,3}^2] = E[I_{1,2,3}^2]$, but $E[W_{1,2,3}^3] \neq E[I_{1,2,3}^3]$ because of the following:

$$E[I_{1,2,3}^3] \leq E(\mu_1 + \mu_2 - X_1 - X_2)^+ = E(X_1 + X_2 - \mu_1 - \mu_2)^+ \leq E[W_{1,2,3}^3]. \quad (5.10)$$

We have assumed that a second surgery would wait to its scheduled starting time if the first surgery were completed early. If we relax that assumption, the second surgery would begin as soon as the first one ends and the probability that the surgery would incur waiting would be the same as the probability that it would incur idleness. However, if operations held to the assumption, the waiting time associated with the second surgery may influence the waiting time associated with that of the third; but any idle time related to the second surgery would not affect the idle time associated with the third. Proposition 17 establishes an exact relationship between waiting- and idle-time associated with the third surgery.

Proposition 17. *For a sequence of three independently distributed surgeries $X_1 \to X_2 \to X_3$, waiting- and idle-times associated with the third surgery satisfy the following relationship:*

$$E[I^3_{1,2,3}] = E[W^3_{1,2,3}] - E[W^2_{1,2,3}]. \tag{5.11}$$

Proof. See the Appendix. □

The objective function for sequence $X_1 \to X_2 \to X_3$, $Z^{\mu_1,\mu_1+\mu_2}_{1,2,3}$, can be formulated as follows:

$$Z^{\mu_1,\mu_1+\mu_2}_{1,2,3} = c^w\{E[W^2_{1,2,3}] + E[W^3_{1,2,3}]\} + c^i\{E[I^2_{1,2,3}] + E[I^3_{1,2,3}]\} + c^o E[O_{1,2,3}]. \tag{5.12}$$

By invoking Propositions 12 and 17, objective function (5.12) can be re-expressed:

$$Z^{\mu_1,\mu_1+\mu_2}_{1,2,3} = c^w E[W^2_{1,2,3}] + (c^w + c^i)E[W^3_{1,2,3}] + c^o E[O_{1,2,3}]. \tag{5.13}$$

Because we have been able to show numerically that $E[O_{1,2}]$ contributes less in determining the optimal sequence than $E[W^2_{1,2}]$ does for the two-surgery case, we anticipate that expected overtime $E[O_{1,2,3}]$ contributes the least in determining the optimal sequence for the three-surgery case.

Further, 441 instances represent the boundary case for which $h = \mu_1 + \mu_2$, the scheduled start time of the third surgery. We observe that conditional expected overtime $E[\bar{O}_{1,2}] := E[O_{1,2}|h = \mu_1 + \mu_2]$, which is the same as $E[W^3_{1,2,3}]$, contributes less in determining the optimal sequence than $E[W^2_{1,2}] = E[W^2_{1,2,3}]$ does. We conclude that $E[W^2_{1,2,3}]$ contributes most in determining the optimal sequence.

In particular, Proposition 18 gives the objective function $Z^{\mu_1,\mu_1+\mu_2}_{1,2,3}$ for three surgeries with *i.i.d.* normal distributions $N(\mu_1, \sigma^2_1), N(\mu_2, \sigma^2_2)$ and $N(\mu_3, \sigma^2_3)$:

Proposition 18. *For a sequence of three surgeries with independent durations that have distributions* $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ *and* $N(\mu_3, \sigma_3^2)$, $Z_{1,2,2}^{\mu_1,\mu_1+\mu_2}$ *can be formulated as :*

$$Z_{1,2,3}^{\mu_1,\mu_1+\mu_2} = c^w \frac{\sigma_1}{\sqrt{2\pi}} + (c^w + c^i)\Big\{\frac{\sigma_1+\sigma_2}{2\sqrt{2\pi}} + \frac{\sigma_1\sigma_2}{2\pi}\Big\} + c^o E[O_{1,2,3}]. \tag{5.14}$$

Proof. See the Appendix. □

Equation (5.14) shows that the SV rule minimizes SWIP when each of three surgery durations follows the normal distribution. The objective function increases with σ_1 and σ_2, and the smaller of σ_1 and σ_2 should be used to designate the first surgery in the sequence because the first term of (5.14) is an increasing function of only σ_1 and the second term of (5.14) is independent of the first two surgeries.

5.5 Application of Results to Scheduling $N = 2k$ Surgeries in k ORs

We demonstrate how our results can be applied by using them as a basis for a heuristic that assigns surgeries to multiple ORs and sequences them in each OR. The typical hospital operates several ORs in each of which two or three surgeries are scheduled. The optimal sequencing rule depends on the first surgery scheduled, which determines SWIP. We conjecture that balancing workloads over ORs will result in a favorable total expected overtime penalty.

Consider assigning two surgeries to each of k ORs with the objectives of minimizing both SWIP and OTP. We deal with multiple ORs, each with a single time block of duration h. Because SWIP dominates OTP, we assign the k surgeries with smallest variance, one to each OR in the first round. After assigning k surgeries according to the SV rule, each to one OR, the second round assigns successively the surgery with the largest mean duration to the OR to which the surgery with the

smallest mean duration has been assigned until two surgeries have been assigned to each OR. The first round seeks favorable SWIP values; and the second, to balance expected workloads over ORs and thus obtain favorable OTP values. Otherwise, some ORs may incur large OTPs, while others incur none. Intuitively, balanced workloads can be expected to be associated with a lower total expected (considering all ORs) overtime than unbalanced workloads.

The following numerical example illustrates our heuristic using $k = 4$ ORs and $2k = 8$ surgeries with each duration following either the lognormal (LN) or normal (N) distribution. Means and variances are given in Table 5.8. The time unit is an hour and $h = 8$; the sum of the expected durations of any three surgeries is greater than h hours so that at most two surgeries are scheduled in each OR to avoid excessive overtime. The number of possible sequences, excluding symmetric instances, is $7 \times 5 \times 3 \times 2^4 = 1,680$, in which $7 \times 5 \times 3$ is the number of assignments to four ORs and 2^4 is the number of possible sequences in the 4 ORs.

Table 5.8: Distributions of Eight Surgery Durations (Time Unit : Hour)

Surgery	1	2	3	4
Distribution	$N(2.5, 0.25^2)$	$N(2.5, 0.5^2)$	$LN(3, 0.3^2)$	$LN(3, 0.6^2)$
Surgery	5	6	7	8
Distribution	$LN(3.5, 0.35^2)$	$LN(3.5, 0.7^2)$	$N(4, 0.4^2)$	$LN(4, 0.8^2)$

Because the SV rule is better than the SM rule w.r.t. SWIP for two surgeries, each with either lognormally or normally distributed durations, we select the $k = 4$ surgeries by the SV rule and assign each to an OR, to obtain favorable $E[W_{1,2}^2]$. Step 1 identifies the first four surgeries as X_1, X_3, X_5 and X_7, and assigns them to OR 1, 2, 3 and 4 (without loss of generality), respectively.

Now, our heuristic assigns a second surgery to each OR. The expected cost would be decreased by assigning a surgery that leads to a low $E[O_{1,2}]$ value for each OR.

Thus, our heuristic assigns the unassigned surgery with the largest mean duration to the OR with the assigned surgery that has the smallest mean duration. Each row in Table 5.9 gives the pair of surgeries assigned to each OR.

Table 5.9: Final Assignment and Sequence

OR	First Surgery(X_f)	Second surgery(X_s)	$E(X_f - \mu_f)^+$	$E[OT_{f,s}]$
$OR1$	$X_1 \sim N(2.5, 0.25^2)$	$X_8 \sim LN(4, 0.8^2)$	0.099	0.000
$OR2$	$X_3 \sim LN(3, 0.3^2)$	$X_6 \sim LN(3.5, 0.7^2)$	0.119	0.018
$OR3$	$X_5 \sim LN(3.5, 0.35^2)$	$X_4 \sim N(3, 0.6^2)$	0.139	0.004
$OR4$	$X_7 \sim N(4, 0.4^2)$	$X_2 \sim N(2.5, 0.5^2)$	0.160	0.002

We have evaluated for all 6 possible pairwise switchings among the four surgeries scheduled second in the ORs and have found that the current sequence in Table 5.9 gives is best. In general, our heuristic prescribes effective schedules because it seeks favorable SWIP first and SWIP dominates OTP if $(c^w + c^i)/c^o$ is not small. Further, it balances expected OR workloads with the goal of obtaining favorable OTP.

5.6 Insights

One of our performance measures, SWIP is closely related to the variance, which is a measure of deviation from the mean. Expected waiting time, $E(X - \mu)^+$, which is the same as expected idle time, $E(\mu - X)^+$, is another measure of deviation from the mean. These partial expectations are equivalent to the mean absolute deviations from the mean: $E(X - \mu)^+ = E(\mu - X)^+ = \frac{1}{2}E|X - \mu|$, which is related to the variance: $E|X - \mu| = K\sqrt{E[(X - \mu)^2]}$, where K is a constant and particular for each distribution (Kenney and Keeping, 1962). For example, it is well known that the ratio of absolute mean deviation to standard deviation is $\sqrt{\frac{2}{\pi}}$ for the normal distribution; that is, $E|X - \mu| = \sigma\sqrt{\frac{2}{\pi}}$ as shown in Section 5.2. Intuitively, a rule that is based on variance is recommendable to minimize SWIP. We observe that the SV rule is better than the SM rule w.r.t. SWIP in the majority of instances and gives equal results in the remaining cases. In particular, we have been able to

show analytically that the SV rule minimizes SWIP for the case of two normally distributed surgeries.

It is important that the first surgery scheduled in a block should be selected judiciously. OTP does not impact the objective function value more than SWIP for the three distributions and two combinations we study. Numerical tests show that, when $h = \mu_1 + \mu_2$ (e.g., instances 9 and 10 in Tables 5.2, 5.3, 5.4, 5.6 and 5.7), $E[\bar{O}_{1,2}]$ contributes less in determining the optimal sequence than $E[W^2_{1,2}]$ does, even though $E[\bar{O}_{1,2}]$ is greater than $E[W^2_{1,2}]$. In particular, sequences $X_1 \to X_2$ and $X_2 \to X_1$ have the same expected overtime if two surgery durations are normally distributed and $h = \mu_1 + \mu_2$. $E[W^3_{1,2,3}]$ contributes less in determining the optimal sequence than $E[W^2_{1,2,3}]$ does for the three-surgery case. The first surgery has a larger impact on objective function values for both two- and three-surgery cases. We conjecture that, as more surgeries are scheduled in a block, surgeries later in the sequence contribute less in determining the optimal sequence than earlier ones, even though surgeries scheduled later contribute more to the expected amount of overtime.

When the exact distribution is not known, we may apply the SV rule to minimize the total expected cost of waiting-, idle- and over-time penalties. For the majority of (but not all) instances we tested, the SV rule is efficacious w.r.t. SWIP for all distributions considered in this paper, expected waiting times $E(X-\mu)^+$ are nearly the same regardless of the distribution of surgery duration, and OTP contributes less in determining the optimal sequence than SWIP does.

6. $D/\tilde{M}/1$ APPOINTMENT SYSTEM

This chapter investigates sequencing rules for a $D/\tilde{M}/1$ appointment system, assuming that customers arrive at deterministic times (denoted D) and are processed according to independent, non-identical exponentially distributed service times (denoted $\tilde{M}$). In particular, it seeks to evaluate the efficacy of two common sequencing rules - smallest-variance-first (SV) and smallest-mean-first (SM) - for cases that involve either two or three customers. The objective is to minimize the sum of expected costs of customer-waiting- and server-idle-times (WIT). The case we study represents systems that, for example, schedule according to time blocks or deal with different classes of customers, each involving few customers. To focus purely on the impact of the sequencing rule, this paper does not consider no-shows or random arrival times.

Queueing models for appointment systems (e.g., $D/M/1$) usually assume that customers arrive for service at pre-determined times rather than randomly (Wang, 1993) and that service times are *i.i.d.* exponential (denoted M). Jansson (1966) studied the $D/M/1$ queueing model to prescribe the optimal inter-arrival time with the objective of minimizing WIT, assuming that the service times of all customers *i.i.d.* exponential. We assume that service times are independent, but not necessarily identical. Weiss (1990) showed that, if surgery times are *i.i.d.* and the distribution is symmetrical, the SV rule minimizes WIT. Gupta (2007) and Denton et al. (2007) used stochastic ordering to sequence two surgeries with durations that have the same mean but different variances with the objective of minimizing WIT.

6.1 Case 1: Two Customers

This section considers two customers with independent, non-identical exponential service times, X_1 and X_2, with means μ_1 and μ_2, respectively. Let t_i be the arrival

time of the i^{th} customer, $i = 1, 2$. Without loss of generality, consider the sequence in which customer 1 precedes customer 2: $X_1 \rightarrow X_2$.

Let $Z_{1,2}^{t_2}$ denote the objective function value for the case in which the first customer is ready at time $t_1 = 0$; and the second customer, at time t_2. We assume that the second customer begins service at time $\max(X_1, t_2)$. The amount of earliness, $(t_2 - X_1)^+$, represents the time during which the server is idle before the second customer arrives. The amount of tardiness, $(X_1 - t_2)^+$, represents the time during which the second customer must wait for service to begin. Our analysis involves costs per unit time for customer waiting, c^w, and server idleness, c^i. The objective function value for sequence $1 \rightarrow 2$, $Z_{1,2}^{t_2}$, is:

$$Z_{1,2}^{t_2} = c^w E[(X_1 - t_2)^+] + c^i E[(t_2 - X_1)^+]. \tag{6.1}$$

We investigate two ways of specifying customer arrival time t_2. The first approach is based on a practical assumption; and the second, on optimizing arrival time.

6.1.1 Optimal Sequence with a Practical Assumption

Objective function (6.1) depends on arrival time, $t2$. Consider the specific assumption that $t_2 = \mu_1$, which is commonly used (Choi and Wilhelm, 2012a; Pinedo, 2009), for example, in scheduling surgeries. By definition of partial expected value, $E[(X_1 - \mu_1)^+] = E[(\mu_1 - X_1)^+]$ (Choi and Wilhelm, 2012a). So, under the assumption that $t_2 = \mu_1$, objective function (6.1) reduces to (6.2):

$$Z_{1,2}^{\mu_1} = (c^w + c^i) E[(X_1 - \mu_1)^+]. \tag{6.2}$$

If service time X_1 is exponentially distributed, objective function (6.2) can be

further specialized:

$$Z_{1,2}^{\mu_1} = (c^w + c^i) \int_{\mu_1}^{\infty} (x_1 - \mu_1) \frac{1}{\mu_1} e^{-\frac{x_1}{\mu_1}} dx_1 = (c^w + c^i) \frac{\mu_1}{e}. \tag{6.3}$$

Objective function, $Z_{1,2}^{\mu_1}$ increases with μ_1. Since the exponential distribution has mean μ_1 and variance μ_1^2, both of which are functions of μ_1, we conclude that both SV and SM rules prescribe the same optimal sequence for two customers under the assumption $t_2 = \mu_1$.

6.1.2 Optimal Sequence with the Optimal Arrival Time

In this subsection, we determine the optimal sequence when the second customer arrives at the optimal time, t_2^*. First, we apply the newsvendor model to prescribe the optimal arrival time:

$$\min_{t_2} \{ Z_{1,2}^{t_2} | t_2 \geq 0 \}.$$

Proposition 19. *$Z_{1,2}^{t_2}$ attains its minimum at $t_2^* = \mu_1 \ln \frac{c^w + c^i}{c^i}$, which increases with μ_1.*

Proof. The optimal solution t_2^* is defined as $t_2^* = F_{X_1}^{-1}(c^w/(c^w + c^i))$, where $F_{X_1}(x)$ is the distribution function of random variable X_1. With X_1 following the exponential distribution, $F_{X_1}(t_2) = 1 - e^{-t_2/\mu_1}$. Combining, we obtain the optimal arrival time, t_2^*:

$$t_2^* = \mu_1 \ln \frac{c^w + c^i}{c^i}. \tag{6.4}$$

To evaluate $Z_{1,2}^t$ for general arrival time t (i.e., without the restriction that $t_2 =$

μ_1), after suppressing subscripts, we obtain $E[(X-t)^+]$ and $E[(t-X)^+]$:

$$E[(X-t)^+] = \int_t^\infty (x-t)\frac{1}{\mu}e^{-\frac{x}{\mu}}dx = \mu e^{-\frac{t}{\mu}} \tag{6.5}$$

$$\text{and } E[(t-X)^+] = \int_0^t (t-x)\frac{1}{\mu}e^{-\frac{x}{\mu}}dx = t-\mu+\mu e^{-\frac{t}{\mu}}. \tag{6.6}$$

Substituting t_2^* as defined in (6.4) for t in (6.5) and (6.6) and, in turn, substituting these expected values in objective function (6.1), we obtain:

$$Z_{1,2}^{t_2^*} = c^w\mu_1 e^{-\frac{t_2^*}{\mu_1}} + c^i\{t_2^* - \mu_1 + \mu_1 e^{-\frac{t_2^*}{\mu_1}}\} = \mu_1 c^i \log\frac{c^w+c^i}{c^i}, \tag{6.7}$$

which is an increasing function of μ_1. □

We conclude that both SV and SM rules prescribe the same optimal sequence for two customers, given that the second one arrives at the optimal time, t_2^*.

6.2 Case 2: Three Customers

Consider three customers with independent, exponentially distributed service times with means, μ_1, μ_2, and μ_3, respectively. Given sequence $X_1 \to X_2 \to X_3$, we assume that the arrival time of the second customer is μ_1 and that of the third customer is $\mu_1+\mu_2$. We do not prescribe optimal arrival times because the required analysis is mathematically intractable. The following two subsections derive a closed form of the objective function and evaluate the optimal sequencing rule, respectively.

6.2.1 Three-customer Objective Function

We evaluate the objective function value of the sequence $X_1 \to X_2 \to X_3$ to prescribe the optimal sequencing rule. The objective function for three customers,

$Z_{1,2,3}^{\mu_1,\mu_1+\mu_2}$, is given by

$$\begin{aligned} Z_{1,2,3}^{\mu_1,\mu_1+\mu_2} &= c^w\{E[(X_1-\mu_1)^+]+E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+]\} \\ &+ c^i\{E[(\mu_1-X_1)^+]+E[(\mu_1+\mu_2-\max(X_1,\mu_1)-X_2)^+]\}. \quad (6.8) \end{aligned}$$

Choi and Wilhelm (2012a) have shown that the expected idle time associated with the third customer is less than his/her expected waiting time. The exact relation is given by (6.9).

$$E[(\mu_1+\mu_2-\max(X_1,\mu_1)-X_2)^+] = E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+]-E[(\mu_1-X_1)^+]. \quad (6.9)$$

Incorporating (6.9), (6.8) can be reduced to the following:

$$Z_{1,2,3}^{\mu_1,\mu_1+\mu_2} = c^w E[(X_1-\mu_1)^+]+(c^w+c^i)E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+].$$

To express $Z_{1,2,3}^{\mu_1,\mu_1+\mu_2}$ in closed form, we must evaluate $E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+]$.

Proposition 20. *The waiting time of the third customer is given by:*

$$E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+] = \begin{cases} \infty & \mu_1 \geq \mu_2 \\ \frac{1}{e^2}(e\mu_2+\mu_1+\frac{\mu_1^2}{\mu_2-\mu_1}) & \textit{otherwise.} \end{cases} \quad (6.10)$$

Proof. $E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+]$ can be decomposed into

$$\begin{aligned} &\int_0^{\mu_1}\int_0^{\infty}(x_2-\mu_2)^+\frac{1}{\mu_1}e^{-\frac{x_1}{\mu_1}}\frac{1}{\mu_2}e^{-\frac{x_2}{\mu_2}}dx_2dx_1 \\ &+\int_{\mu_1}^{\infty}\int_0^{\infty}(x_1+x_2-\mu_1-\mu_2)^+\frac{1}{\mu_1}e^{-\frac{x_1}{\mu_1}}\frac{1}{\mu_2}e^{-\frac{x_2}{\mu_2}}dx_2dx_1. \quad (6.11) \end{aligned}$$

If $\mu_1 > \mu_2$, second integral term in (6.11) is unbounded. Otherwise, (6.11) reduces to a closed form:

$$E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+] = \frac{e-1}{e^2}\mu_2+\frac{\mu_2^2}{\mu_2-\mu_1}e^{-2} = \frac{1}{e^2}\Big(e\mu_2+\mu_1+\frac{\mu_1^2}{\mu_2-\mu_1}\Big). \ \square$$

Hence, objective function value, $Z_{1,2,3}^{\mu_1,\mu_1+\mu_2}$, is given by (6.12) if $\mu_1 < \mu_2$:

$$Z_{1,2,3}^{\mu_1,\mu_1+\mu_2} \quad = \quad c^w\frac{\mu_1}{e} + (c^w+c^i)\frac{1}{e^2}\big[e\mu_2+\mu_1+\frac{\mu_1^2}{\mu_2-\mu_1}\big]. \tag{6.12}$$

6.2.2 Optimal Sequencing Rule

We can determine the first customer to be the one with the smallest variance (or mean), because if $\mu_1 > \mu_2$, $Z_{1,2,3}^{\mu_1,\mu_1+\mu_2}$ goes to infinity. To determine an optimal order of customers in sequence positions, define, term (6.12), $f_{\mu_1}(\mu_2) := [e\mu_2+\mu_1+\mu_1^2/(\mu_2-\mu_1)]$ and fix the value of μ_1. $e\mu_2+\mu_1$ is an affine function with slope e and y-intersect μ_1, and $\mu_1^2/(\mu_2-\mu_1)$ is a convex function of μ_2.

Figure 6.1 graphs $f_{\mu_1}(\mu_2)$ over the range $-\infty < \mu_2 < \infty$. We focus on the upper-right curve of Figure 6.1 (a), which represents the range of $\mu_2 \geq \mu_1$. Other figures (b)-(e) depict selected subregions over the range of $\mu_2 > \mu_1$. $f_{\mu_1}(\mu_2)$ can be shown to be a convex function of μ_2 that attains its minimum at $(1+\frac{1}{\sqrt{e}})\mu_1 \simeq 1.607\mu_1$.

We compare two sequences: $X_1 \to X_2 \to X_3$ and $X_1 \to X_3 \to X_2$, accordingly, $f_{\mu_1}(\mu_2)$ and $f_{\mu_1}(\mu_3)$. If $\mu_1 \leq \mu_2 \leq 1.607\mu_1$ and $\mu_1 \leq \mu_3 \leq 1.607\mu_1$ (Subregion I in Figure 6.1 (b)) , $f_{\mu_1}(\mu_2)$ is a decreasing function of μ_2. If $\mu_2 > 1.607\mu_1$ and $\mu_3 > 1.067\mu_1$ (Subregion II in Figure 6.1 (c)), $f_{\mu_1}(\mu_2)$ is an increasing function of μ_2. For any $t > (\sqrt{e}+1)^2\mu_1$, let α_t and β_t be the solutions to $t = e\mu_2+\mu_1+\mu_1^2/(\mu_2-\mu_1)$ such that $\alpha_t < \beta_t$ ((Subregion III and IV in Figures 6.1 (d) and (e), respectively).

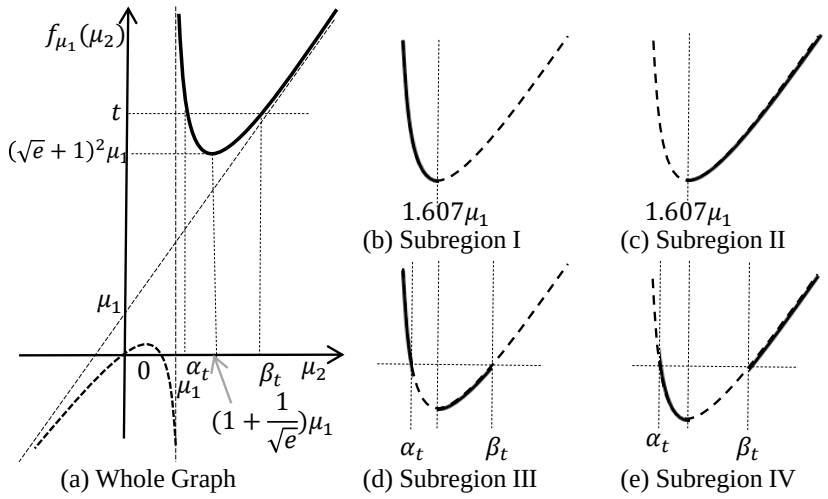


Figure 6.1: Graph of $f_{\mu_1}(\mu_2)$ and Separate Regions.

The optimal sequence can be summarized as follows:

$$\begin{cases} 1 \to 3 \to 2 & \text{if } \mu_1 < \mu_2, \mu_3 < 1.607\mu_1, \quad \mu_2 < \mu_3 \quad (SubregionI) \\ 1 \to 2 \to 3 & \text{if } \mu_1 < \mu_2, \mu_3 < 1.607\mu_1, \quad \mu_2 > \mu_3 \quad (SubregionI) \\ 1 \to 2 \to 3 & \text{if } \mu_2, \mu_3 > 1.607\mu_1, \quad \mu_2 < \mu_3 \quad (SubregionII) \\ 1 \to 3 \to 2 & \text{if } \mu_2, \mu_3 > 1.607\mu_1, \quad \mu_2 > \mu_3 \quad (SubregionII) \\ 1 \to 3 \to 2 & \text{if } \mu_1 < \mu_2 < \alpha_t, \quad \beta_t < \mu_3 \quad (SubregionIII) \\ 1 \to 2 \to 3 & \text{if } \alpha_t < \mu_2 < 1.607\mu_1, \quad \mu_3 > \beta_t \quad (SubregionIV). \end{cases} \tag{6.13}$$

Neither SV or SM prescribes the optimal sequence in this case. Given that the sequencing rule for the three-customer case is so complex, we conclude that a single criterion like SV or SM cannot prescribe the optimal sequence for the general case. However, the customer with the smallest parameter (implementing both SV and SM rules) must be sequenced in the first position.

7. CONCLUSION

This chapter summarizes all findings from analytical and numerical studies at each level, and provides several venues for the future research direction.

7.1 Summary

We present a prototypical model to optimize the allocation of surgical specialties to OR days and four adaptations (i.e., models), along with associated solution approaches to facilitate solution: (NV-CA) news vendor-based capacity allocation, (NV-SIP) news vendor-based stochastic integer programming, (SIP) stochastic integer programming, and (NS-SIP) stochastic integer programming without symmetry. It reports numerical tests that compare the computational characteristics of the models.

We obtain solutions with less variability within a few seconds using NV-CA and NV-SIP. Hence, we recommend that NV-SIP be used to support detailed allocation decisions; and NV-CA, for rough-cut capacity planning. The NV-CA solution could provide a better framework for MSS planning if it were disaggregated into the allocation of individual specialties to OR days.

Comparing the run times required to resolve the ($|N|$, $|M|$) levels (5,5) and (5,5)x(5,5) shows that it is better to decompose a problem into components $\kappa \in \mathcal{K}$ for solution. ($|N|$, $|M|$) levels (5,5)x(5,5) and (10,10) both deal with ten specialties and ten ORs, but the latter allows any of the specialties to be allocated to any of the ORs. As to be expected, this flexibility allows somewhat better solutions to be found, especially for larger surgery-duration CV. However, the run time required to determine solutions with the same level of precision increases because more allocation alternatives must be investigated.

We present new methods to prescribe optimal planned duration and sequence of time blocks, each of which reserves OR resources for a particular surgical subspecialty at the tactical level. Further, rather than using an overbooking policy, it gives a closed form to prescribe optimal planned block duration to hedge no shows. Results lend considerable insights for managing OR resources.

The methods we propose for MSS can be implemented easily and, we expect, would result in improved performance through managing the MSS process and optimizing the sum of expected earliness and lateness costs. Effectively planned block durations can also be expected to facilitate scheduling of actual patients at the operational level.

We confirm the efficacy of the SV rule to sequence surgeries in each time block at the operational level. We examine rules for sequencing two surgeries with durations that follow either the lognormal, gamma, or normal distribution. We are able to obtain a closed form of $E[W_{1,2}^2]$ for each of the three cases and to conclude analytically that the SV rule is optimal if both surgeries follow the normal distribution. We show numerically that the SV rule is better in determining the optimal sequence than the SM rule for the majority of our test instances and that the two rules give the same result in remaining instances.

We show numerically that lognormal, gamma and normal distributions all give very similar values of $E[W_{1,2}^2]$ and $E[I_{1,2}^2]$. Thus one may pick the most tractable distribution when we do not know the exact form. We study sequencing two surgeries for cases in which the lognormal distribution is used in combination with either the gamma or normal distribution. Numerical tests show that the SV rule is better than the SM rule w.r.t. SWIP. If $(c^w + c^i)/c^o$ is not small, $\Delta E[O]$ does not determine the sequencing of surgeries, even when expected overtime is greater than expected waiting time. We recommend that the SV rule be used to obtain favorable SWIP

and OTP values.

We study the three-surgery case in which all durations are normally distributed, conducting numerical tests to evaluate $\Delta E[O]$ and $\Delta E[W]$. The expected waiting time associated with a third surgery has a lesser effect on determining the optimal sequence than that associated with the second surgery. We conclude that scheduling the first surgery is the most important and advocate use of the SV rule in making this assignment.

To demonstrate how our results might be applied, we describe how they can be used as the basis for a heuristic to assign surgeries to multiple ORs and sequence them, assuming that only two surgeries can be accommodated in a time block. Because SWIP contributes more in determining the optimal sequence than OTP does, the first surgery in each OR is more important in determining the optimal sequence than the second, which contributes only to OTP, not SWIP.

7.2 Future Works

We suggest several avenues for future research. Future research could integrate capacity allocation and expansion (e.g., addition of a new OR) decisions over with a longer planning horizon. Another fruitful direction would develop improved algorithms to solve SIP and NS-SIP. Future research could also could devise a superior means for breaking the symmetry of model SIP, perhaps by including tighter constraints.

Our findings suggest several avenues for future research at the tactical level. For example, an MSS may affect staff scheduling, PACU, and other relevant departments. Incorporating such ancillary departments in MSS planning is an opportunity for the future research. Future research could fruitfully address the multi-OR problem. Finally, our model of the *sequential newsvendor* problem can be applied in

time-sensitive environments other than health care (e.g., JIT delivery) because both earliness and lateness must be minimized at the same time.

This work opens several avenues for future research in sequencing surgeries. We assume that all patients arrive punctually, but this may not be possible in reality, so modeling random patient arrivals provides an opportunity for future research. Further, we do not considered no-shows. We use the sum of expected durations of the previous surgeries as the ready time of the next patient. This has not been shown to be optimal; but it facilitates analysis, follows prior research, and provides a rule that can be followed easily in practice. Future research could optimize patient ready times along with other performance measures such as waiting-, idle-, and over-time penalties.

REFERENCES

Adan, I., J. Vissers. 2002. Patient mix optimisation in hospital admission planning: A case study. *International Journal of Operations and Production Management* **22** 445 – 461.

Bai, G., S. Hsu, R. Krishnan. 2009. Accounting performance, cost structure, and firm's capacity investment decisions. *Retrieved from http://jindal.utdallas.edu/files/Paper-Krishnan.pdf* 17 Aug 2009.

Bazaraa, M. S., H. D. Sherali, C. M. Shetty. 2006. *Nonlinear Programming Theory and Algorithms*. 3rd ed. Wiley Interscience, New Jersey.

Beliën, J., E. Demuelemeester. 2006. Building cycle master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research* **176** 1185–1204.

Beliën, J., E. Demuelemeester. 2008. A branch-and-price approach for integrating nurse and surgery scheduling. *European Journal of Operational Research* **189** 652 – 668.

Beliën, J., E. Demuelemeester, B. Cardoen. 2009. A decision support system for cyclic master surgery scheduling with multiple objectives. *Journal of Scheduling* **12** 147–161.

Birge, J. R., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer, New York.

Blake, J., J. Donald. 2002. Mount sinai hospital uses integer programming to allocate operating room time. *Interface* **32** 63 – 73.

Blake, J. T. 2011. Capacity planning in operating rooms. Y. Yih, ed., *Handbook of healthcare delivery systems*. CRC Press, Boca Raton, 34–1 – 34–12.

Blake, J. T., M. W. Carter. 1997. Surgical process scheduling: A structured review. *Journal of Health Systems* **5** 17–30.

Blake, J. T., F. Dexter, J. Donald. 2002. Operating room manager's use of integer programming for assigning block time to surgical groups: A case study. *Anesthesia and Analgesia* **94** 143–148.

Bruno, J., P. Downey, G. N. Frederickson. 1981. Sequencing tasks with exponential service times to minimize the expected flow time or makespan. *Journal of the ACM* **28** 100–113.

Cardeon, B., E. Demeulemeester, J. Beliën. 2009. Sequencing surgical cases in a day-care environment: An exact brance-and-price approach. *Computers & Operations Research* **36** 2660 – 2669.

Cardoen, B., E. Demeulemeester, J. Beliën. 2010. Operating room planning and scheduling: A literature review. *European Journal of Operational Research* **201** 921 – 932.

Carr, S., W. Lovejoy. 2000. The inverse newsvendor problem: Choosing an optimal demand portfolio for capacitated resources. *Management Science* **46** 912–927.

Casella, G., R. G. Berger. 2001. *Statistical Inference*. 2nd ed. Thomson Learning, Australia.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: a review of literature. *Production and Operations Management* **12** 519 – 549.

Chakraborty, S., K. Muthuraman, M. Lawley. 2010. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions* **42** 354 – 366.

Choi, S., M. Ketzenberg. 2012. An inverse newsvendor model to set the optimal number of customers in a capacitated environment. *Working paper* .

Choi, S., W. E. Wilhelm. 2012a. An analysis of sequencing surgeries with durations that follow the lognormal, gamma, or normal distribution. *IIE Transactions on Healthcare Systems and Engineering* **2** 156–171.

Choi, S., W. E. Wilhelm. 2012b. An approach to optimize master surgical block schedules (submitted to). *Operations Research* .

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* **35** 1003 – 1016.

Denton, B., J. Viapiano, A. Vogl. 2007. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science* **10** 13 – 24.

Denton, B. T., A. J. Miller, H. J. Balasubramanian, T. R. Huschka. 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research* **58** 1 – 15.

Dexter, F., M. Hopwood. 1999. An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. *Anesthesia and Analgesia* 7–20.

Dexter, F., J. Ledolter, R. Wachtel. 2002. Sampling error can significantly affect measured hospital financial performance of surgeons and resulting operating room time allocation. *Anesthesia and Analgesia* **95** 184–188.

Dexter, F., J. Ledolter, R. Watchel. 2005. Tactical decision making for selective expansion of operating room resources incorporating financial criteria and uncertainty in subspecialties' future workloads. *Anesthesia and Analgesia* **100** 1425 – 1432.

Dexter, F., A. Macario, R. D. Traub, D. A. Lubarsky. 2003. Operating room utilization alone is not an accurate metric for the allocating of operating room block time

to individual surgeons with low caseloads. *Anesthesia and Analgesia* **98** 1243–1249.

Dexter, F., L. O'Neill. 2004. Data envelopment analysis to determine by how much hospitals can increase elective inpatient surgical workload for each specialty. *Anesthesia and Analgesia* **99** 1492 – 1500.

Fei, H., C. Chu, N. Meskens. 2009. Solving a tactical operating room planning problem by a column-generation-based heuristic procedure with four criteria. *Annals of Operations Research* **166** 91 – 108.

Fei, H., C. Chu, N. Meskens, A. Artiba. 2008. Solving surgical cases assignment problem by a branch-and-price approach. *International Journal of Production Economics* **112** 96 – 108. Special Section on Recent Developments in the Design, Control, Planning and Scheduling of Productive Systems.

Fei, H., N. Meskens, C. Chu. 2010. A planning and scheduling problem for an operating theatre using an open scheduling strategy. *Computers and Industrial Engineering* **58** 221 – 230.

Folland, G. B. 1999. *Real analysis: modern techniques and their applications*. 2nd ed. John & Sons Wiley, New York.

Glazebrook, K. D. 1979. Scheduling tasks with exponential service times on parallel processors. *Journal of Applied Probability* **16** 685–689.

Green, L. V. 2004. Capacity planning and management in hospitals. M.L. Brandeau, F. Sainfort, W.P. Pierskalla, eds., *Operations Research and Healthcare: A Handbook of Methods and Applications*. Springer, New York, 15–41.

Guerriero, F., R. Guido. 2010. Operational research in the management of the operating theatre: a survey. *Health Care Management Science* **14** 89 – 114.

Guinet, A., S. Chaabane. 2003. Operating theatre planning. *International Journal of Production Economics* **85** 69 – 81.

Gul, S., B. T. Denton, J. W. Fowler, T. Huschka. 2011. Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management* **20** 406 – 417.

Gupta, D. 2007. Sugical suites' operations management. *Production and Operations Management* **16** 689 – 700.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: challenges and opportunities. *IIE Transactions* **40** 800 – 819.

Gupta, D., L. Wang. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research* **56** 567–592.

Guse, C. E., L. Richardson, M. Carle, K. Schmidt. 2003. The effect of exit-interview patient education on no-show rates at a family practice residency clinic. *The Journal of the American Board of Family Practice* **16** 399 – 404.

Higle, J. L. 2005. Stochastic programming: Optimization when uncertainty matters. *Tutorials in Operations Research* 30–53.

Jansson, B. 1966. Choosing a good appointment system: A study of queues of the type(d/m/1). *Operations Research* **14** 292–312.

Jebali, A., A. B. Hajalouane, P. Ladet. 2006. Operating rooms scheduling. *International Journal of Production Economics* **99** 52 – 62.

Kaandorp, G. C., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Science* **10** 217 – 229.

Kanich, D., J. Byrd. 1996. How to increase efficiency in operating room. *Surgical Clinics of North America* **76** 161 – 173.

Kenney, J. F., E. S. Keeping. 1962. *Mathematics of Statistics*. 3rd ed. Van Nostrand, New York.

Kharraja, S., P. Albert, S. Chaabane. 2006. Block scheduling: Toward a master surgi-

cal schedule. *International Conference on Service Systems and Service Management* 429 – 435.

Kuo, P. C., R. A. Schroeder, S. Mahaffey, R. R. Bollinger. 2003. Optimization of operating room allocation using linear programming techniques. *Journal of American College of Surgeons* **197** 889 – 895.

Lamiri, M., F. Grimaud, X. Xie. 2009. Optimization methods for a stochastic surgery planning problem. *International Journal of Production Economics* **120** 400 – 410.

Lamiri, M., X. Xie, A. Dolgui, F. Grimaud. 2008. A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research* **185** 1026 – 1037.

Lebowitz, P. 2003. Schedule the short procedure first to improve or efficiency. *AORN Journal* **78** 651 – 659.

Lin, J., K. Muthuraman, M. Lawley. 2011. Optimal and approximate algorithms for sequential clinical scheduling with no-shows. *IIE Transactions on Healthcare Systems Engineering* **1** 20–36.

Lovejoy, W. S., Y. Li. 2002. Hospital operating room capacity expansion. *Management Science* **48** 1369 – 1387.

May, J., D. Strum, L. Vargas. 2000. Fitting the lognormal distribution to surgical procedure times. *Decision Sciences* **31** 129 – 148.

May, J. H., W. E. Spangler, D. P. Strum, L. G. Vargas. 2011. The surgical scheduling problem: current research and future opportunities. *Production and Operations Management* **20** 392 – 405.

Nahmias, S. 2008. *Production and Operations Analysis*. 6th ed. McGraw-Hill/Irwin, New York.

Niño-Mora, J. 2002. Stochastic scheduling. F.A. Christodoulos, P. Pardalos, eds.,

Encyclopedia of Optimization. Kluwer, Dordrecht, 367–370.

Olivares, M., C. Terwiesch, L. Cassorla. 2008. Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Science* **54** 41–55.

Pinedo, M. L. 2008. *Scheduling: Theory, Algorithms and Systems*. 3rd ed. Springer, New York.

Pinedo, M. L. 2009. Planning and scheduling in health care. *Planning and Scheduling in Manufacturing and Services*. Springer, New York, 291 – 316.

Porteus, E. 2002. *Foundations of Stochastic Inventory Theory*. Stanford University Press, Stanford.

Righter, R. 1994. Stochastic scheduling. Moshe Shaked, ed., *Stochastic orders and their applications*. Academic Press, Boston, 381 – 432.

Rohleder, T. R., D. Sabapathy, R. Shorn. 2005. An operating room block allocation model to improve hospital patient flow. *Clinical and Investigative Medicine* **28** 353 – 355.

Samanlioglu, F., Z. Ayag, B. Batili, E. Evcimen, G. Yilmaz, O. Atalay. 2010. Deterministic master schedule of surgical operations by integer programming a case study. *Proceedings of the 2010 Industrial Engineering Research Conference* .

Santibanez, P., M. Begen, D. Atkins. 2007. Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a british columbia health authority. *Health Care Management Science* **10** 269 – 282.

Sier, D., P. Tobin, C. McGurk. 1997. Scheduling surgical procedure. *Journal of the Operational Research Society* **48** 884 – 891.

Stepaniak, P. S., C. Heij, G. de Veries. 2009. Modeling and prediction of surgical procedure times. *Retrieved from http://ideas.repec.org/p/dgr/eureir/1765017017.html*

11 Mar 2009.

Strum, D., J. May, A. Sampson, L. Vargas, W. Spangler. 2003. Estimating times of surgeries with two components procedures. *Anesthesiology* **98** 232 – 240.

Strum, D., J. May, L. Vargas. 2000a. Modeling the uncertainty of surgical procedure times: comparison of the log-normal and normal models. *Anesthesiology* **92** 1160 – 1167.

Strum, D., A. Sampson, J. May, L. Vargas. 2000b. Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology* **92** 1454 – 1466.

Strum, D., L. Vargas, G. May, G. Bashein. 1997. Surgical suite utilization and capacity planning: A minimal cost analysis model. *Journal of Medical Systems* **21** 309– 322.

Testi, A., E. Tanfani, G. Torre. 2007. A three-phase approach for operating theatre schedules. *Health Care Management Science* **10** 163 – 172.

Tiwari, V., D. H. Berger. 2010. Elective surgery scheduling. *INFORMS 2010 Annual Meeting*. Austin.

van Oostrum, J. M., M. Van Houdenhoven, J. L. Hurink, E. W. Hans, G. Wullink, G. Kazemier. 2008. A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum* **30** 355 – 374.

Wachtel, R. E., F. Dexter. 2008. Tactical increases in operating room block time for capacity planning should not be based on utilization. *Anesthesia and Analgesia* **106** 215 – 226.

Wachtel, R. E., F. Dexter. 2010. Review of behavioral operations experimental studies of newsvendor problems for operating room management. *Anesthesia and Analgesia* **110** 1698 – 1710.

Wang, P. P. 1993. Static and dynamic scheduling of customer arrivals to a single-

server system. *Naval Research Logistics* **40** 345 – 360.

Wang, P. P. 1997. Optimally scheduling N customer arrival times for a single-server system. *Computers & Operations Research* **24** 703 – 716.

Weber, R. R. 1982. Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime. *Journal of Applied Probability* **19** 167 – 182.

Weber, R. R., P. Varaiya, J. Walrand. 1986. Scheduling jobs with stochastically ordered processing times on parallel machines to minimize expected flowtime. *Journal of Applied Probability* **23** 841 – 847.

Weiss, E. N. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions* **22** 143 – 150.

Wikipedia. 2010. Log normal distribution. *Retrieved from http:en.wikipedia.orgwikiLog-normal_distribution* 7 May 2010.

Zhang, B., M. M. Dessouky, D. Belson. 2009. A mixed integer programming approach for allocating operating room capacity. *Journal of the Operational Research Society* **60** 663 – 673.

APPENDIX A

PROOFS

Proof of Proposition 6

Proof.

$$
\begin{aligned}
E[(y-T)^+] &= \int_{-\infty}^{y}(y-t)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \\
&= (y-\mu)\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt + \int_{-\infty}^{y}(\mu-t)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \\
&= (y-\mu)\Phi(\frac{y-\mu}{\sigma}) + \frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}
\end{aligned}
$$

□

Proof of Proposition 7

Proof.

$$
\begin{aligned}
E[(T-y)^+] &= \int_{y}^{\infty}(t-y)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \\
&= \int_{y}^{\infty}(t-\mu)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt + (\mu-y)\int_{y}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \\
&= [-\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}]_{y}^{\infty} + (\mu-y)\int_{y}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \\
&= \frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}} + (\mu-y)[1-\Phi(\frac{y-\mu}{\sigma})].
\end{aligned}
$$

□

Proof of Proposition 8

Proof. Let $\hat{y}$ be the solution that minimizes $f(y)$, and z be the standard normal score $z = \frac{\hat{y}-\mu}{\sigma}$ at the optimal solution, $\Phi(z) = \Phi(\frac{\hat{y}-\mu}{\sigma}) = \frac{1}{1+\beta} = \frac{c^l}{c^e+c^l}$, and $\hat{y} = \mu + z\sigma$.

$$\begin{aligned}
f(\hat{y}) &= c^e\Big\{\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(\hat{y}-\mu)^2}{\sigma^2}} + (\hat{y}-\mu)\Phi(\frac{\hat{y}-\mu}{\sigma})\Big\} \\
&\quad + c^l\Big\{\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(\hat{y}-\mu)^2}{\sigma^2}} + (\mu-\hat{y})(1-\Phi(\frac{\hat{y}-\mu}{\sigma}))\Big\} \\
&= c^e\Big\{\frac{\sigma}{\sqrt{2\pi}}e^{-z^2} + \sigma z\Phi(z)\Big\} + c^l\Big\{\frac{\sigma}{\sqrt{2\pi}}e^{-z^2} - \sigma z(1-\Phi(z))\Big\} \\
&= (c^e+c^l)\frac{\sigma}{\sqrt{2\pi}}e^{-z^2} + \sigma z\Phi(z)(c^e+c^l) - \sigma z c^l \\
&= (c^e+c^l)\frac{\sigma}{\sqrt{2\pi}}e^{-z^2},
\end{aligned}$$

which is an increasing function of σ. □

Proof of Proposition 12

Proof.

$$\begin{aligned}
E[(X_1-\mu_1)^+] =& \int_{\mu_1}^{\infty}(x-\mu_1)f(x_1)dx_1 \\
=& \int_{-\infty}^{\infty}(x_1-\mu_1)f(x_1)dx_1 - \int_{-\infty}^{\mu_1}(x_1-\mu_1)f(x_1)dx_1 \\
=& \int_{-\infty}^{\mu_1}(\mu_1-x_1)f(x_1)dx_1 = E[(\mu_1-X_1)^+].
\end{aligned}$$

□

Proof of Proposition 13

Proof. The probability density function, the cumulative distribution function, the expected value and the partial expected value of the Lognormal distribution are well known as follows: (Wikipedia, 2010):

$$\begin{aligned} f(x_1 : \mu_1, \sigma_1) &= \frac{1}{\sqrt{2\pi}\sigma_1 x_1} e^{-\frac{(\log x_1 - \mu_1)^2}{2\sigma_1^2}} \\ F(x_1 : \mu_1, \sigma_1) &= \Phi\Big(\frac{\log x_1 - \mu_1}{\sigma_1}\Big) \\ E(X_1) &= e^{\mu_1 + \frac{1}{2}\sigma_1^2} \\ \int_k^\infty x_1 f(x_1) dx_1 &= e^{\mu_1 + \frac{1}{2}\sigma_1^2} \Phi\Big(\frac{\mu_1 + \sigma_1^2 - \log k}{\sigma_1}\Big). \end{aligned}$$

Substituting these definitions into $E[(X_1 - E(X_1))^+]$ gives the following:

$$\begin{aligned} E[(X_1 - E(X_1))^+] =& \int_{E(X_1)}^\infty (x_1 - E(X_1)) f(x_1) dx_1 \\ =& \int_{E(X_1)}^\infty x_1 f(x_1) dx_1 - \int_{E(X_1)}^\infty E(X_1) f(x_1) dx_1 \\ =& E(X_1)\Phi\Big(\frac{\mu_1 + \sigma_1^2 - (\mu_1 + \frac{1}{2}\sigma_1^2)}{\sigma_1}\Big) \\ & - E(X_1)\Big(1 - \Phi(\frac{\mu_1 + \frac{1}{2}\sigma_1^2 - \mu_1}{\sigma_1})\Big) \\ =& E(X_1)\Big(2\Phi(\frac{\sigma_1}{2}) - 1\Big) = e^{\mu_1 + \frac{1}{2}\sigma_1^2}\Big(2\Phi(\frac{\sigma_1}{2}) - 1\Big) \end{aligned}$$

□

Proof of Proposition 14

Proof. We note that the expected value $E(X_1)$ and c.d.f. $F(x_1)$ of the gamma distribution are given as follows:

$$\begin{aligned} E(X_1) =& \mu_1 = n\beta \\ F(x_1) =& \int_0^{x_1} \frac{x^{n-1}e^{-\frac{x}{\beta}}}{\beta^n \Gamma(n)} dx = 1 - \sum_{i=0}^{n-1} \frac{(x_1/\beta)^i}{i!} e^{-\frac{x_1}{\beta}}. \end{aligned}$$

The partial expected value $E[(X_1 - \mu_1)^+]$ is given as follows:

$$\begin{aligned} E[(X_1 - \mu_1)^+] =& \int_{\mu_1}^{\infty} (x_1 - \mu_1) x_1^{n-1} \frac{e^{-\frac{x_1}{\beta}}}{\beta^n \Gamma(n)} dx_1 \\ =& \int_{\mu_1}^{\infty} \frac{x_1^n e^{-\frac{x_1}{\beta}}}{\beta^n \Gamma(n)} dx_1 - \mu_1 \int_{\mu_1}^{\infty} \frac{x_1^{n-1} e^{-\frac{x_1}{\beta}}}{\beta^n \Gamma(n)} dx_1 \\ =& n\beta \sum_{i=0}^{n} \frac{(\mu_1/\beta)^i}{i!} e^{-\frac{\mu_1}{\beta}} - n\beta \sum_{i=0}^{n-1} \frac{(\mu_1/\beta)^i}{i!} e^{-\frac{\mu_1}{\beta}} \\ =& n\beta \frac{n^n}{n!} e^{-n}. \end{aligned}$$

□

Proof of Proposition 15

Proof.

$$\begin{aligned} E[(X_1 - \mu_1)^+] &= \int_{\mu_1}^{\infty} (x_1 - \mu_1) \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} dx_1 \\ &= \left[-\frac{\sigma_1}{\sqrt{2\pi}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \right]_{\mu_1}^{\infty} = \frac{\sigma_1}{\sqrt{2\pi}}. \end{aligned}$$

□

Proof of Proposition 16

Proof.

$$
\begin{aligned}
&E[(\max(X_1,\mu_1)+X_2-d)^+]\\
&=\int_{-\infty}^{\mu_1}\int_{-\infty}^{\infty}(\mu_1+x_2-d)^+ f_{X_1}(x_1)f_{X_2}(x_2)dx_2dx_1\\
&+\int_{\mu_1}^{\infty}\int_{-\infty}^{\infty}(x_1+x_2-d)^+ f_{X_1}(x_1)f_{X_2}(x_2)dx_2dx_1\\
&=\int_{-\infty}^{\mu_1}\int_{d-\mu_1}^{\infty}(\mu_1+x_2-d) f_{X_1}(x_1)f_{X_2}(x_2)dx_2dx_1\\
&+\int_{\mu_1}^{\infty}\int_{d-x_1}^{\infty}(x_1+x_2-d) f_{X_1}(x_1)f_{X_2}(x_2)dx_2dx_1\\
&=\int_{-\infty}^{\mu_1} f_{X_1}(x_1)dx_1\int_{d-\mu_1}^{\infty}(\mu_1+x_2-d) f_{X_2}(x_2)dx_2\\
&+\int_{\mu_1}^{\infty}\int_{(d-x_1)}^{\infty}(x_1+x_2-d) f_{X_1}(x_1)f_{X_2}(x_2)dx_2dx_1\\
&=\frac{1}{2}E[(\mu_1+X_2-d)^+]+\int_{\mu_1}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}\frac{\sigma_2}{\sqrt{2\pi}}e^{-\frac{(x_1-d+\mu_2)^2}{2\sigma_2^2}}dx_1\\
&\qquad+\int_{\mu_1}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}(x_1-d+\mu_2)(1-\Phi(\frac{d-x_1-\mu_2}{\sigma_2}))dx_1.
\end{aligned}
$$

Now, consider two cases as follow:

(i) $X_1+X_2<h$ a.s.

$$P(X_1+X_2>h)=0 \text{ a.s. and } P(\mu_1+X_2>h)=0 \text{ a.s.}$$

$$E[(\max(X_1,\mu_1)+X_2-h)^+]\approx 0.$$

(ii) $|X_1 + X_2 - h| < \epsilon$, let $X_1 + \mu_2 = h$ and $\mu_1 + X_2 = h$.

$$\int_{\mu_1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \frac{\sigma_2}{\sqrt{2\pi}} e^{-\frac{(x_1-d+\mu_2)^2}{2\sigma_2^2}} dx_1 = \int_{\mu_1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \frac{\sigma_2}{\sqrt{2\pi}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_2^2}} dx_1$$

$$= \frac{\sigma_1\sigma_2^3}{2\pi(\sigma_1^2+\sigma_2^2)}.$$

The remaining term is simplified as follows:

$$\int_{\mu_1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} (x_1 - d + \mu_2)\Big(1 - \Phi(\frac{d - x_1 - \mu_2}{\sigma_2})\Big) dx_1$$

$$= \int_{\mu_1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} (x_1 - \mu_1)\Big(1 - \Phi(\frac{\mu_1 - x_1}{\sigma_2})\Big) dx_1$$

$$= \Big[-\frac{\sigma_1}{\sqrt{2\pi}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \Big(1 - \Phi(\frac{\mu_1 - x_1}{\sigma_2})\Big)\Big]_{\mu_1}^{\infty} + \int_{\mu_1}^{\infty} \frac{\sigma_1}{\sqrt{2\pi}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_2^2}} dx_1$$

$$= \frac{\sigma_1}{2\sqrt{2\pi}} + \frac{\sigma_1^3\sigma_2}{2\pi(\sigma_1^2+\sigma_2^2)}.$$

Thus, $E[(\max(X_1, \mu_1) + X_2 - d)^+] = \frac{\sigma_1+\sigma_2}{2\sqrt{2\pi}} + \frac{\sigma_1\sigma_2^3+\sigma_1^3\sigma_2}{(2\pi)(\sigma_1^2+\sigma_2^2)} = \frac{\sigma_1+\sigma_2}{2\sqrt{2\pi}} + \frac{\sigma_1\sigma_2}{2\pi}$. □

Proof of Proposition 17

Proof.

$$
\begin{aligned}
&E[(\mu_1+\mu_2-\max(X_1,\mu_1)-X_2)^+]\\
&=\int_{-\infty}^{\mu_1}\int_{-\infty}^{\infty}(\mu_2-x_2)^+f_{X_2}(x_2)f_{X_1}(x_1)dx_2dx_1\\
&+\int_{\mu_1}^{\infty}\int_{-\infty}^{\infty}(\mu_1+\mu_2-x_1-x_2)^+f_{X_2}(x_2)f_{X_1}(x_1)dx_2dx_1\\
&=\int_{-\infty}^{\mu_1}E[(\mu_2-X_2)^+]f_{X_1}(x_1)dx_1\\
&+\int_{\mu_1}^{\infty}E[(\mu_1+\mu_2-X_1-X_2)^+]f_{X_1}(x_1)dx_2dx_1\\
&=\int_{-\infty}^{\mu_1}E[(\mu_2-X_2)^+]f_{X_1}(x_1)dx_1+\int_{\mu_1}^{\infty}E[(\mu_1+\mu_2-X_1-X_2)^+]f_{X_1}(x_1)dx_1\\
&=\int_{-\infty}^{\mu_1}E[(X_2-\mu_2)^+]f_{X_1}(x_1)dx_1\\
&+\int_{\mu_1}^{\infty}\{E[(x_1+X_2-\mu_1-\mu_2)^+]-E[x_1+X_2-\mu_1-\mu_2]\}f_{X_1}(x_1)dx_1\\
&=\int_{-\infty}^{\mu_1}E[(X_2-\mu_2)^+]f_{X_1}(x_1)dx_1+\int_{\mu_1}^{\infty}E[(x_1+X_2-\mu_1-\mu_2)^+]f_{X_1}(x_1)dx_1\\
&-\int_{\mu_1}^{\infty}(x_1-\mu_1)f_{X_1}(x_1)dx_1\\
&=E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+]-E[(X_1-\mu_1)^+].
\end{aligned}
$$

□

Proof of Proposition 18

Proof.

$$E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+]$$

$$=\int_{-\infty}^{\mu_1}\int_{-\infty}^{\infty}(x_2-\mu_2)^+f_{X_2}(x_2)f_{X_1}(x_1)dx_2dx_1+$$

$$\int_{\mu_1}^{\infty}\int_{-\infty}^{\infty}(x_1+x_2-\mu_1-\mu_2)^+f_{X_2}(x_2)f_{X_1}(x_1)dx_2dx_1$$

$$=\frac{1}{2}E[(X_2-\mu_2)^+]+\int_{\mu_1}^{\infty}\int_{\mu_1+\mu_2-x_1}^{\infty}(x_1+x_2-\mu_1-\mu_2)f_{X_2}(x_2)f_{X_1}(x_1)dx_2dx_1$$

$$=\frac{\sigma_2}{2\sqrt{2\pi}}+\int_{\mu_1}^{\infty}E(X_2+x_1-\mu_1-\mu_2)^+f_{X_1}(x_1)dx_1$$

$$=\frac{\sigma_2}{2\sqrt{2\pi}}+\int_{\mu_1}^{\infty}\frac{\sigma_2}{\sqrt{2\pi}}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_2^2}}f_{X_1}(x_1)dx_1+\int_{\mu_1}^{\infty}(x_1-\mu_1)\Phi(\frac{x_1-\mu_1}{\sigma_2})f_{X_1}(x_1)dx_1.$$

The second term of the above will be as follows:

$$\int_{\mu_1}^{\infty}\frac{\sigma_2}{\sqrt{2\pi}}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_2^2}}f_{X_1}(x_1)dx_1=\int_{\mu_1}^{\infty}\frac{\sigma_2}{\sqrt{2\pi}}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_2^2}}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}dx_1$$

$$=\frac{\sigma_1\sigma_2^3}{2\pi(\sigma_1^2+\sigma_2^2)}.$$

The last term will be as follows:

$$\begin{aligned}
\int_{\mu_1}^{\infty}(x_1-\mu_1)\Phi(\frac{x_1-\mu_1}{\sigma_2})f_{X_1}(x_1)dx_1 &= \int_{\mu_1}^{\infty}(x_1-\mu_1)\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}\Phi(\frac{x_1-\mu_1}{\sigma_2})dx_1 \\
&= \Big[-\frac{\sigma_1}{\sqrt{2\pi}}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}\Phi(\frac{x_1-\mu_1}{\sigma_2})\Big]_{\mu_1}^{\infty} \\
&\quad + \int_{\mu_1}^{\infty}\frac{\sigma_1}{\sqrt{2\pi}}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_2^2}}dx_1 \\
&= \frac{\sigma_1}{2\sqrt{2\pi}} + \frac{\sigma_1^3\sigma_2}{2\pi(\sigma_1^2+\sigma_2^2)}
\end{aligned}$$

Thus, $E[(\max(X_1,\mu_1)+X_2-\mu_1-\mu_2)^+] = \frac{\sigma_1+\sigma_2}{2\sqrt{2\pi}} + \frac{\sigma_1\sigma_2}{2\pi}$, which is increasing in σ_2 at fixed σ_1. □

APPENDIX B

SUMMARY OF NOTATION

Notation for Chapter 3

Index Sets and Indices

M	ORs	$m \in M$
N	surgery specialties, (e.g., orthopedic, cardiovascular)	$n \in N$
$\mathcal{K}$	compatible surgical specialties and ORs	$\kappa \in \mathcal{K}$
M_κ	ORs dedicated to specialties $n \in N_\kappa$	
N_κ	specialties to be performed in ORs $m \in M_\kappa$	
D	days (e.g., Monday through Friday)	$d \in D = \{1, \ldots, 5\}$

Parameters

P_n	random, representative duration of each surgery
A_n	forecast number of surgeries demanded each period
h	standard OR-day
c_n^a	penalties for each surgery of specialty $n \in N$ that is not accommodated
c_n^u	penalties for under-usage of OR time relative to h
c_n^o	penalties for over-usage of OR time relative to h

Decision Variables

R_n	the number of OR days to which specialty $n \in N_\kappa$ is allocated
V_n	the number of representative surgeries requiring specialty n that are assigned each day to each OR in set M_κ to which specialty n is allocated

Random Variables

U_n	$:= \max(h - [V_n * P_n], 0)$, under-usage of each OR day
O_n	$:= \max([V_n * P_n] - h, 0)$, over-usage of each OR day
$\bar{A}_n$	$:= \max(A_n - R_n V_n, 0)$, the number of patients requiring specialty n who are not accommodated
$\bar{S}_n$	$:= \min(A_n, R_n V_n)$, the number of patients requiring specialty n who are accommodated

Notation for Chapter 4

Index Sets and Indices

I	sub-specialties	$i \in I$
K	sequence positions for time blocks	$k \in K$
Δ	permutations of time blocks	$\delta \in \Delta$

Parameters

c^e	Earliness penalty cost
c^l	Lateness penalty cost
β	Ratio of earliness cost to lateness cost, $\beta = c^e/c^l$
$B^{\delta}_{[k]}$	random block duration of kth block under sequence $\delta \in \Delta$
$T^{\delta}_{[k]}$	random block end time of kth block under sequence $\delta \in \Delta$

Decision Variables

$x^{\delta}_{[k]}$	planned block duration of kth block under sequence $\delta \in \Delta$
$y^{\delta}_{[k]}$	planned end time of kth block under sequence $\delta \in \Delta$

Notation for Chapter 5

Index Sets and Indices

J	patients	$j \in J$

Parameters

c^w	waiting time penalty cost
c^i	idle time penalty cost
c^o	overtime penalty cost
h	time block duration
t_j	ready-time associated with patient $j \in J$
X_j	surgery duration for patient $j \in J$
$W^2_{1,2}$	$:= (X_1 - t_2)^+$, waiting time associated with the second surgery
$I^2_{1,2}$	$:= (t_2 - X_1)^+$, idle time associated with the second
$O_{1,2}$	$:= [\max(X_1, t_2) + X_2 - h]^+$, overtime
$Z^{t_2}_{1,2}$	objective function value for the case in which the sequence of surgeries is 1,2