Statistical Analyses of Call Center Data

Research Thesis

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Polyna Khudyakov

Submitted to the Senate of the Technion - Israel Institute of Technology

Tammuz, 5770 Haifa July, 2010

This Research Thesis Was Done Under the Supervision of
Professor Malka Gorfine and Professor Paul D. Feigin
in the Faculty of Industrial Engineering.

The Generous Financial Help of the Technion
is Gratefully Acknowledged.

Abstract

This study looks into management problems of call centers and the opportunity to analyze a large quantity of data collected over a long time period. The aim is to develop and apply methods of statistical analysis to call center data in order to identify basic problems, to find the sources of such problems, to develop ways for their solution and to estimate their possible impact.

We consider Markovian models for a call center with and without an Interactive Voice Response (IVR) system and approximate performance in the Quality and Efficiency Driven (QED) asymptotic regime, which is suitable for moderate to large call centers. In contrast to exact calculations, the approximations are both insightful and easy to implement (for up to thousands of agents). We validate our models against data from a US Bank Call Center, and our results demonstrate that simple models still provide very useful descriptions of much more complex realities.

We also present a statistical analysis of customers patience. This work is the first attempt to apply frailty models to an analysis of customers' patience while taking into account the possible dependency between calls of the same customer, and estimating this dependency.

We extended the estimation technique of Gorfine et al. [37] to address the case of different unspecified baseline hazard functions for each call, to address the case in which customer's behavior changes as s/he becomes more experienced with the call center services. Then, we provided a new class of test statistics for hypothesis testing of the equality of the baseline hazard functions. The asymptotic distribution of the test statistics was investigated theoretically under the null hypothesis and certain local alternatives. We also provided variance estimator. The properties of the test statistics, under finite sample size, were studied by an extensive simulation study and verified the control of Type I error and our proposed sample size formula. The utility of our proposed estimating technique is illustrated by the analysis of the call center data of an Israeli commercial company that processes up to 100,000 calls per day. According to this analysis, customers are more patient in their first call. The differences between customers' patience in the second, third and fourth calls are not significant.

Key words: Queues, Closed Queueing Networks; Call or Contact Centers, Impatience, Busy Signals; IVR, VRU; QED or Halfin-Whitt regime; Asymptotic Analysis; Multivariate Survival Analysis, Frailty Model, Customer Patience, Hypothesis Testing, Nonparametric Baseline Hazard Function.

Contents

List of Acronyms

IVR Interactive Voice Response

ACD Automatic Call Distributor

IEEE Institute of Electrical and Electronics Engineers

MCMC Markov Chain Monte Carlo

PASTA Poisson Arrivals See Time Averages

QED Quality and Efficiency Driven

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Our Goals

In our increasingly industrialized and globalized world, a large number of companies include call centers in their structures and more than $300 billion is spent annually on call centers around the world [34]. For a customer, addressing the call center actually means addressing the company itself, and any negative experience on the part of the customer can lead to the rejection of company products and services. Hence, for the company, it is very important to ensure that a call center functions effectively and provides high quality service to its customers.

Call centers collect a huge amount of data, and this provides a great opportunity for companies to use this information for the analysis of customer needs, desires, and intentions. Such data analysis can help improve the quality of customer service and lower the costs. A typical call center spends about two-thirds of its operational costs on salaries. However, it would be a false economy to reduce costs by decreasing the number of agents, because a small change in staffing level can have a dramatic impact upon the level of service. Thus, a major goal of a call center manager is to establish an appropriate tradeoff between its expenses and its service level. We propose queueing models that can help reach sound decisions by yielding performance-analysis tools that support this tradeoff. We also supplement our theory with statistical analysis of our model's primitive - customer patience.

1.2 An Analysis of Call Center Performance

In order to achieve high-quality customer service and effective management of operating costs, many leading companies are deploying new technologies, such as enhanced Interactive Voice Response (IVR) devices, natural speech self-service options and others. IVR systems are specialized technologies designed to enable self-service of callers, without the assistance of human agents. The IVR technology helps call centers to keep costs from rising (and sometimes to reduce costs), while hopefully improving service levels, revenue and hence profits.

Our work develops and analyzes models, for a call center with and without an IVR. We find analytical formulae which describe typical call center performance measures, such as the probability of a busy signal, the probability of abandonment and the average waiting time for an agent. The use of these formulae helps us to analyze the impact of different parameters on the operational system performance and to find the relationship between the number of agents and other system parameters depending on the desired level of service.We also provide an empirical study in order to evaluate the value of adding an IVR, which is based on analyzing real data from a large call center.

1.3 Customer Patience Analysis

One of our models' primitives is a customer patience, which we define as the ability to endure waiting for service. This human trait plays an important role in the call center mechanism. As mentioned above, every call can be considered as a possibility to keep or to lose a customer, and the outcome depends on the customer's satisfaction. Moreover, customers are likely to remember one disappointing service experience more clearly than twenty good ones. From this point of view, an abandoned call is a negative experience which affects the future customer's choice.

There are different factors affecting the customer's waiting behavior. Only some of them are observable and available to us, and these are included in the model as covariates. Unobservable factors that are likely to influence the customer's patience are different customer's characteristics and customer's temperament. In this work, we use a model that takes into account observed and unobserved personal customer's features; and this provides a great advance in customer patience analysis. In addition, we investigate the effect of the customers'

experiences on their waiting behavior.

1.4 The Structure of the Work

Chapter 2 contains a survey of the literature dealing with related works. In Section 2.2 we review the literature concerning mathematical models of a call center and analysis of operational performance measures. Literature related to customer patience analysis is considered in Section 2.3.

Chapter 3 deals with the design and analysis of theoretical models describing a typical call center. In Section 3.1 we consider the extension of the model proposed by Srinivasan et al. [80] by assuming finite customer patience and the M/M/S/N+M queue model. Then, in Section 3.2 we find approximations for frequently used performance measures, which support decision-making for call center managers and help in the analysis of the staffing problem. An analysis of the accuracy of the approximations is presented in Section 3.3. A detailed comparison between exact and approximated performance shows that the approximations often work perfectly, even outside the Quality and Efficiency Driven (QED) regime. Section 3.4 summarizes our findings through practical rules-of-thumb (expressed via the offered load) and we chart the boundary of this "outside". In Section 3.5, we validate our approximations against data from a real call center, thus establishing their applicability. For the convenience of the reader, the proofs of theorems from Section 3.2 are presented in Section 3.6. In Section 3.7 we summarize our findings and propose future directions for research.

Customer patience is analyzed in **Chapter 4**. In Section 4.1 we start with a description of the data that motivated the study. In Section 4.2 we briefly explain the choice of our model. Section 4.3 presents the notation and formulation of the model. The estimating procedure and the asymptotic properties of the estimators are presented in Section 4.4. A new test for comparing of two or more baseline hazard functions in the case of dependent observations is provided in Section 4.5. In Section 4.6 we propose a sample size formula for given significance level and power. The proofs and technical details are presented in separate section, namely in Section 4.7. The utility of our proposed estimating technique, a test for comparison and a sample size formula are illustrated in Section 4.8 by extensive simulation study. Then, in Section 4.9, we apply the results of our approach to the real call center data. Our conclusions and future work are set out in Section 4.10. Although our research was motivated by call center data, the

proposed methods can also be of practical importance in different research fields. Thus, in Section 4.10.1, we apply our approach for analyzing breast cancer data of family study.

In **Chapter 5** we summarize the results of our work and discuss the innovation proposed in our study, the methodology used and possible scientific and practical contributions.

Chapter 2

LITERATURE REVIEW

2.1 Descriptive Statistical Analysis

Statistical analysis of call center data started with the creation of call centers. The work of Roberts [75], Duffy and Mercer [23], Liu [59] and Kort [55] written in the 1970s are dedicated mostly to the description and analysis of models with customer abandonments and retrials that took place as a result of telephone network impairments. The underlying research was initiated by companies providing telephone services and telephone equipment.

The study by Liu [59] can be considered as a continuation of the survey conducted in [23]. Liu's main goal was to provide a comprehensive characterization of network performance and customer behavior in setting up a customer's desired telephone connection. Using the collected data, Liu summarized various statistical characteristics, i.e. initial attempts at disposition probabilities, retrial probabilities, the number of additional attempts, ultimate success probabilities and distribution functions for retrial intervals following different types of uncompleted initial attempts.

Kort [55] described models and methods developed at Bell Laboratories to evaluate customer acceptance of telephone connections in the Bell System Public Switched Telephone Network. The models that were developed and used in this study provided a basis for IEEE standards for telephone network performance specifications in a multi-vendor environment. The detailed description of data analyzed in our work can be found in Donin et al. [22] and Trofimov et al. [83].

2.2 An Analysis of a Call Center Performance

A detailed survey of literature on queuing models for call center design are provided by Gans et al. [30].

2.2.1 The QED Regime

The mathematical framework considered here is a multi-server heavy-traffic asymptotic regime, which is referred to as the QED (Quality and Efficiency Driven) regime. Systems that operate in the QED regime enjoy a combination of very high efficiency together with very high quality of service, as surveyed by Gans et al. [30]. A mathematical characterization of the QED regime for the GI/M/S queue was established by Halfin and Whitt [38] as having a non-trivial limit (within (0,1)) of the fraction of delayed customers, with S increasing indefinitely. The latter characterization was also established for GI/D/S (Jelenkovic et al. [47]), M/M/S with exponential patience (Garnett et al. [31]) and with general patience (Mandelbaum and Zeltyn [63]).

The QED regime was explicitly recognized as early as 1923 in Erlang's paper (that appeared in [27]), which addresses both Erlang-B (M/M/S/S) and Erlang-C (M/M/S) models. Later extensive related work took place in various telecom companies but little has been publicly documented. A precise characterization of the asymptotic expansion of the blocking probability, for Erlang-B in the QED regime, was given by Jagerman [46], Whitt [86], and then Massey and Wallace [65] for the analysis of finite buffers. The phenomenon of abandonment in a call center with multiple servers was analyzed by Garnett et al. [31] (Erlang-A model (M/M/S+M)) and Mandelbaum and Zeltyn [63] (M/M/S+G).

2.2.2 The Square-Root Staffing Principle

Erlang's characterization of the QED regime was in terms of the *square-root staffing principle* (sometimes called the "safety-staffing principle"). The square-root principle has two parts to it: first, the conceptual observation that the safety staffing level is proportional to the square-root of the offered load; and second, the explicit calculation of the proportionality coefficient. Borst et al. [12] developed a framework that accommodates both of these needs. More important, however, is the fact that their approach and framework allow an arbitrary cost structure, having the potential to generalize beyond Erlang-C. The square-

root staffing principle arises also in [65] for the M/M/S/N queue, in [31] for M/M/S+M, and others, as surveyed in Gans et al. [30].

2.2.3 Analytical Models of Call Center Performance

In the detailed introduction to call centers by Gans et al. [30], it is explained how call centers can be modeled by queueing systems of various characteristics. Many results and models with references are surveyed in that paper. The authors examine models of single type customers and single skill agents; models with busy signals and abandonment; skills-based routing; call blending and multi-media; and geographically dispersed call centers.

Figure 2.1 depicts a schematic model of a simple inbound call center with S agents serving one class of customers. A call at either the IVR or within the servers' pool occupies a trunk line. There are N trunk lines in this call center. As shown, the waiting room is limited to $N - S$ waiting positions and waiting customers may leave the system due to impatience. A blocked or abandoning customer might try to call again later (retrial). A queueing model of such an inbound call center is characterized by customer profiles, agent characteristics, queue discipline, and system capacity.

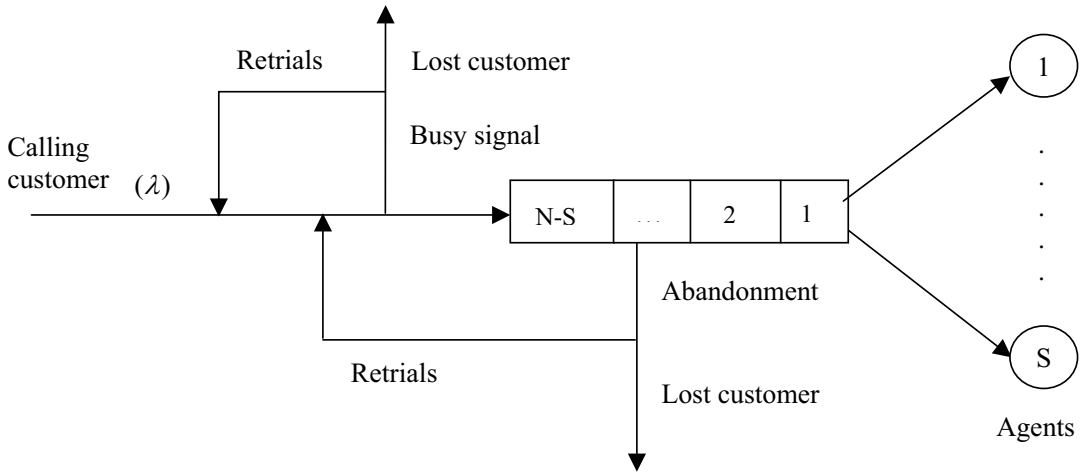


Figure 2.1: Schematic model of a call center with one class of impatient customers, busy signals, retrials and identical agents.

The simplest case with homogeneous customers and homogeneous agents is analytically tractable only if one assumes Poisson arrivals, exponential service times and no retrials. With these assumptions, the underlying stochastic processes are one-dimensional Markov processes, i.e., the future behavior is conditionally independent of the past, given the present state.

The basic operational questions in the design of call centers are: "How can one provide an acceptable quality of service with minimal costs?", or "How many agents and trunk lines do we need in order to provide a given service level?". In general:"How does one balance quality of service with operational efficiency?"

Frequently used measures which support decision-making include the average length of waiting time in the queue, the probability of encountering a busy signal, the probability of waiting, agents' occupancy, etc. In order to analyze the staffing problem, analytical models have been developed in order to help find the answer. The most widely-used model is M/M/S, which is also known as Erlang-C. In this model, the arrival process is Poisson, the service time distribution is exponential and there are S independent, statistically identical agents. It is the simplest yet most prevalent model that supports call center staffing.

The M/M/S model allows an unlimited number of customers in the system but, in practice, this number is limited by the number of trunk lines. This gives rise to the model M/M/S/N (when $S = N$, it is called Erlang-B). Massey and Wallace [65] proposed a procedure for determining the appropriate number of agents S and telephone trunk lines N needed by call centers. They constructed a new efficient search method for the optimal S and $N - S$ that satisfies a given set of Service Level Agreement (SLA) metrics. Moreover, they developed a second approximate algorithm using steady-state, QED-based asymptotic analysis that in practice is much faster than the search method. The asymptotically derived number of agents and the number of waiting spaces in the buffer are found by iteratively solving a fixed point equation.

There are several possibilities to model a call center and the choice of an appropriate model depends on the problem to be solved and the possibility of finding a solution. Generally, most convenient models for such an analysis are of an open type, i.e. they do not have restrictions on the number of places in the system. Such models were considered previously (Mandelbaum et al. [62], Aguir et al. [6], Harris et al. [39]). However, in some cases, it is reasonable to use a closed model, i.e. a model with a limited number of places. For instance, de Vericourt and Jennings [84] dealt with the problem of hospital staffing when they

had to take into consideration the number of places in the system, namely, an always finite number of beds in a given hospital. Another type of closed model was considered by Randhawa and Kumar [73]. Their system was limited to a number of subscribers. As mentioned above, such a model is appropriate for communication systems.

Analytical models of a Call Center with an IVR were developed by Brandt et al. [13]. They showed, and we shall use this fact later on, that it is possible to replace the semi-open network of their model with a closed Jackson network. Such a network has the well-known product form solution for its stationary distribution. This product-form distribution was used by Srinivasan et al. [80] in order to calculate expressions for the probability to find all lines busy and the conditional distribution function of the waiting time before service. However, due to the complex nature of these expressions and the numerical instability associated with the computation process, the whole procedure may be time-consuming and ultimately produce inaccurate values. On the other hand, it is possible to use approximations for the system characteristics as was shown in my M.Sc. thesis [50]. These approximations are convenient for the investigation of the effect of changes in the system parameters on the system performance. At the same time, in [50] approximations of a real call center by models with and without an IVR are analyzed, though it did not support possible customer abandonments. In the current work we extend the model presented in [50] by equipping customers with finite patience.

2.3 Customer Patience Analysis

The first model for customer patience was constructed by Palm [68] in 1943. He introduced a so-called time-dependent inconvenience function that is actually a proportional hazard rate function. An important result, postulated by Palm, is the presence of a correlation between a hazard rate of the customer patience time and his/her irritation caused by waiting. Palm also suggested that patience was characterized by a Weibull distribution, a specific case of this distribution being an exponential distribution widely used in queuing theory (Erlang-A queue model). We also use the assumption of exponentially distributed patience time to create a theoretical model of a typical call center (Sections 3.1.1 and 3.1.2). The assumption of Weibull distributed patience also was proposed by Kort [55] who studied customer acceptance of telephone connections. A detailed survey of

the above and other literature on models and methods used for the analysis of customer patience was provided by Gans et al. [30].

A descriptive model of customer patience with the use of real call center data was presented by Brown et al. [16]. They estimated the distribution of customer patience using the standard Kaplan-Meier product limit estimator. The survival functions were created for different types of service. The authors found that customers performing stock trading are willing to wait more than customers calling for regular services. This unexpected result was explained by the fact that these customers need the service more urgently, and have more trust in the system to provide it. In addition, Brown et al. [16] constructed nonparametric hazard rate estimates. Namely, for each interval of length δ, the estimate of the hazard rate was calculated as $\Big[\sharp \text{ of events during } (t, t+\delta]\Big] / \Big[(\sharp \text{ at risk at } t) \times \delta\Big]$. The resulting function had two peaks and these peaks occurred after a "Please wait" message played by the system with 60 seconds difference. This example illustrates that sometimes ostensibly correct management solutions have the opposite effect.

2.3.1 Survival Analysis

The complication of customer patience analysis is that in most cases customers receive the required service before they lose their patience and we do not observe the values of customer patience. We call such incomplete data as censored observations. To analyze the data with censored observations we need tools of survival analysis. Generally, survival analysis involves the modeling of time to event data. The occurrences of these events are often referred to as failures. Failure time data occur in numerous fields including medicine, economics and industry. The basic models of survival analysis are described in Kalbfleisch and Prentice [48], Hougaard [42] and references therein, among others.

2.3.2 Frailty Models

The Cox proportional hazard model is one of the most widely used event history models. It was proposed by Cox [21] and assumes that event times are independent. Thus, for the analysis of correlated (clustered) failure times an extended Cox model was proposed (Ripatti and Palmgren [74], Murphy [66], Parner [70]), in which a random effect, for each cluster, is included in the model. This random effect model is known as frailty model. Frailty model provides a natural approach to account for risk heterogeneity. The cluster-specific random variate

acts multiplicatively on the hazard function. Under a frailty model, the regression coefficients are cluster-specific log-hazard ratios. It is clear that the frailty model is modeling the conditional hazard function given the latent frailty (Hsu et al. [43], Hsu et al. [45], Duffy et al. [24]). This is in contrast with the marginal modeling (Hsu and Gorfine [44], Shih and Chatterjee [78], Lin [58]), where the correlation is modeled through a multivariate distribution which often involves a copula function, with a specified model for the marginal hazard function. The regression coefficients in the marginal model represent the log-hazard ratios at the population level, regardless of which cluster an individual comes from. In our context, when the objective is to make inference about calls of the same customer, a customer-specific risk estimate is more relevant than a population-averaged risk estimate. Zeger et al. [88] provides a comprehensive comparison of cluster-level modelling versus the marginal population-average approach.

Many frailty models have been considered, including Gamma (Klein [51], Nielsen et al. [67]), Positive stable (Hougaard [41]), Inverse Gaussian (Aalen and Gjessing [3]), compound Poisson (Henderson et al. [40], Aalen [2]), and Log-normal (Ripatti and Palmgren [74]). Hougaard [42] provided a broad review of models consists of different frailty distributions. The most commonly used frailty distribution is the Gamma frailty distribution, because of mathematical convenience. However, it is of concern that misspecification of Gamma frailty distribution may invalidate the inference. Different frailty distributions induce different dependence structure, then, it is important to examine the adequacy of the Gamma frailty model for describing the intracluster dependence. Model diagnostic procedures have been developed for that purpose (Shih [77], Glidden [35], Chen et al. [19]). There are also some works dealing with the misspesification of frailty distribution (Glidden and Vittinghoff [36], Kosorok et al. [53]). Hsu et al. [45] studied how the misspecification affects the estimation of the marginal parameters. They analyzed the simulated data under the assumption of Gamma distributed frailty, while the true distributions were Inverse Gaussian, Positive Stable and a specific case of Discrete distribution. This analysis showed that the Gamma distribution appears to be robust to frailty distribution misspecification in cohort and case-control family studies.

A detailed review of methods for estimation and the model testing were provided by Hougaard [42]. Nielsen et al. [67] and Klein [51] considered the NPMLE estimate of the proportional hazard model with gamma frailty. Murphy [66] showed the consistency and asymptotic normality for this model without covari-

ates. Later, Parner [70] extended these results to the model with covariates. Zeng and Lin [90] presented an estimation technique for the class of semiparametric regression models for censored data, which also include the random effects for dependent time failures. They provided a semi-parametric maximum likelihood estimator, based on the EM algorithm, together with their asymptotic properties. A noniterative estimation procedure for estimating the parameters of the frailty model with any frailty distribution with finite moments was proposed by Gorfine et al. [45]. The detailed proof of the asymptotic properties of the proposed estimators was provided by Zucker et al. [91].

2.3.3 Testing for Equality of Hazard Functions

The most popular test statistic for testing the equatlity of two hazard functions is the weighted log-rank test. It was first proposed by Mantel [64] and later Peto and Peto [71] named it log-rank. An adaptation of this test to censored data was suggested by Prentice [72]. Different extensions of the Wilcoxon rank-sum statistic to censored failure time data were also considered (Gehan [32], Peto and Peto [71], and Tarone and Ware [82]). These proposed models together with the log-rank statistic can be incorporated into the class of weighted log-rank statistics. The asymptotic properties of the weighted log-rank statistics were derived via martingale theory (Gill [33], Fleming and Harrington [28], Andersen et al. [8]). The family of log-rank statistics presented by Harrington and Fleming [28] describes a large variety of weighted log-rank statistics such as the log-rank, Prentice-Wilcoxon, Gehan-Wilcoxon and Tarone-Ware statistics.

Often weighted log-rank statistics considered data generated from independent samples (Lawless and Nadeau [56], Cook et al. [20], Eng and Kosorok [26]). Comparison of two treatments based on clustered data with no covariates is presented by Gangnon and Kosorok [29]. They used the weighted log-rank test statistic and presented a simple sample size formula. Song et al. [79] studied a covariate-adjusted weighted log-rank statistic for recurrent events data while comparing between two independent treatment groups. For the best of our knowledge, so far there is no published work that deals with correlated samples test applied to a covariate adjusted frailty model.

2.3.4 Sample Size Formula

One of the most widely used sample size formula for the log-rank test under the setting of two independent samples is that of Schoenfeld [76]. This formula was developed under the assumption that the hazard functions are not time varying. Combining the idea of Schoenfeld and extending the class of alternatives presented by Fleming and Harrington [28], Kosorok and Lin [54] proposed a class of contiguous alternatives for power and sample size calculations. This class was used for sample size calculations for clustered survival data, with no covariates, using the log-rank statistic (Gangnon and Kosorok [29]), for the supremum log-rank statistic (Eng and Kosorok [26]) and for covariate adjusted log-rank statistic for independent samples (Song et al. [79]). In all the above works, the sample size formula was done under simplifying assumptions, such as assuming identical censoring distributions, consistent difference between the two hazard functions, and continuous hazard functions.

Chapter 3

DESIGN AND INFERENCE FOR A TYPICAL CALL CENTER

3.1 Notation and Formulation of Our Models

As mentioned earlier, a call center typically consists of telephone trunk lines, a switching machine known as the Automatic Call Distributor (ACD), an interactive voice response (IVR) unit, and agents to handle the incoming calls. In this chapter we provide theoretical analyses of two models of a typical call center. The first model does not take into account IVR processes and describes only agents' service and waiting before this service. The second model is more complicated and considers a pool of agents together with the service process in the IVR unit.

3.1.1 Call Center without an IVR

We assume that the arrival process is a Poisson process with rate λ. There are N trunk lines in the system, i.e. arriving customers enter the system only if there is an idle trunk line. We assume that customers have finite patience. Under our assumptions, if a call waits in the queue, it may leave the system after an exponentially distributed time, or is answered by an agent, whichever happens first. The rate of abandonments equals δ. Agents' service times are taken to be independent identically distributed exponential random variables with the rate of μ.

In queueing theory the described model is called the M/M/S/N+M queueing

model and schematically can be described as follows:

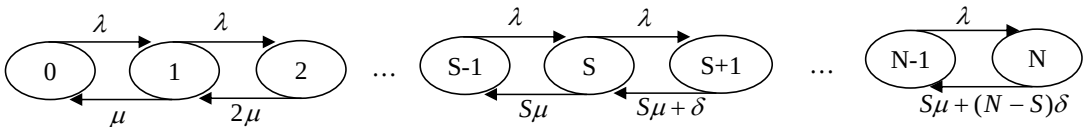


Figure 3.1: M/M/S/N+M queue model.

The M/M/S/N+M queue has the following stationary distribution:

$$\pi_i = \begin{cases} \pi_0 \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i, & 0 \le i \le S; \\ \pi_0 \frac{1}{S!}\left(\frac{\lambda}{\mu}\right)^S \prod_{j=1}^{i-S} \frac{\lambda}{S\mu + j\delta}, & S < i \le N; \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

where

$$\pi_0 = \Big[\sum_{i=0}^{S} \frac{1}{i!}\Big(\frac{\lambda}{\mu}\Big)^i + \sum_{i=S+1}^{N} \frac{1}{S!}\Big(\frac{\lambda}{\mu}\Big)^S \prod_{j=1}^{i-S} \frac{\lambda}{S\mu + j\delta} \Big]^{-1}. \tag{3.2}$$

According to the PASTA theorem [87] we can easily formulate the expressions for operational performance measures. Let W be the waiting time - the time spent by customers, who opt for service, from just after they leave the IVR until being served by an agent. Thus,

- the probability $P(W > 0)$ that a customer waits after the IVR:

$$P(W > 0) = \sum_{i=S}^{N-1} \pi_i, \tag{3.3}$$

- the *probability of abandonment*, given waiting:

$$P(Ab|W > 0) = \frac{\sum_{i=S+1}^{N} \pi_i (S\mu + (i-S)\delta) \frac{(i-S)\delta}{S\mu + (i-S)\delta}}{\sum_{i=S+1}^{N} \pi_i (S\mu + (i-S))}, \tag{3.4}$$

- the *expectation of the waiting time*, given waiting can be calculated using the following relationship:

$$E[W|W > 0] = \frac{P(Ab|W > 0)}{\delta}, \tag{3.5}$$

- the probability to find the system busy (block):

$$P(block) = \frac{\frac{1}{S!}\left(\frac{\lambda}{\mu}\right)^S \prod_{j=1}^{N-S} \frac{\lambda}{S\mu + j\delta}}{\sum_{i=0}^{S} \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i + \sum_{i=S+1}^{N} \frac{1}{S!}\left(\frac{\lambda}{\mu}\right)^S \prod_{j=1}^{i-S} \frac{\lambda}{S\mu + j\delta}}. \tag{3.6}$$

3.1.2 Call Center with an IVR

Now we consider the following model of a call center, as depicted in Figure 3.2: The arrival process is a Poisson process with rate λ. There are N trunk lines and S agents in the system ($S \leq N$). If this is the case, the customer is first served by an IVR processor. We assume that the IVR processing times are independent and identically distributed exponential random variables with rate θ. After finishing the IVR process, a call may leave the system with probability $1 - p$ or proceed to request service from an agent with probability p.

Customer patience is exponentially distributed with rate δ. Agents' service times are taken to be independent identically distributed exponential random variables with rate μ, which are independent of the arrival times and IVR processing times. If a call finds the system full, i.e. all N trunk lines are busy, it is lost (which amounts to a busy signal).

We now view our model as a system with two multi-server queues connected in series (Figure 3.3). The first one represents the IVR processor. This processor can handle at most N jobs at a time, where N is the total number of trunk lines available. The second queue represents the agents' pool which can handle at most S incoming calls at a time. The number of agents is naturally less than or equal the number of trunk lines available, i.e. $S \leq N$. Moreover, N is also an upper bound for the total number of customers in the system: at the IVR plus waiting to be served plus being served by the agents.

Let $Q(t) = (Q_1(t), Q_2(t))$ represent the number of calls at the IVR processor and at the agents' pool at time t, respectively. Since there are only N trunk lines, then $Q_1(t) + Q_2(t) \leq N$, for all $t \geq 0$. Note that the stochastic process $Q = \{Q(t),\ t \geq 0\}$ is a finite-state continuous-time Markov chain. We shall denote its states by the pairs $\{(i, j) \mid i + j \leq N,\ \ i, j \geq 0\}$.

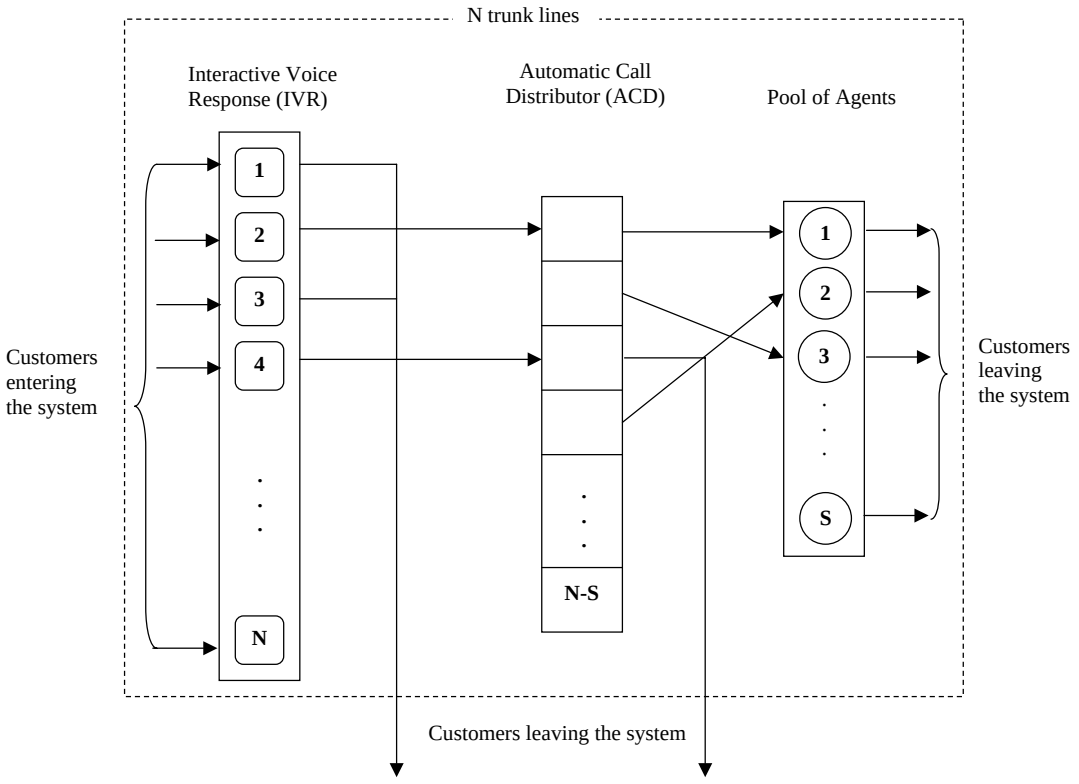


Figure 3.2: Schematic model of a call center with an IVR, S agents, N trunk lines and customers' abandonment.

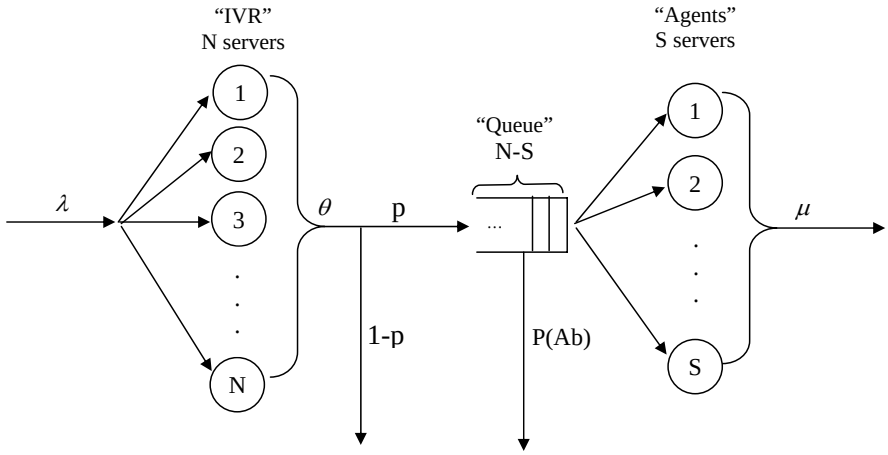


Figure 3.3: Schematic model of a call center with an interactive voice response, S agents and N trunk lines.

As shown in [13], one can consider our model as a 2 stations within a 3-station closed Jackson network, by introducing a fictitious state-dependent queue. There are N entities circulating in the network. Service times in the first, second, and third stations are exponential with rates θ, μ and λ respectively, and the numbers of servers are N, S, and 1, respectively. This 3-station closed Jackson network has a product form solution for its stationary distribution (see Figure 3.7).

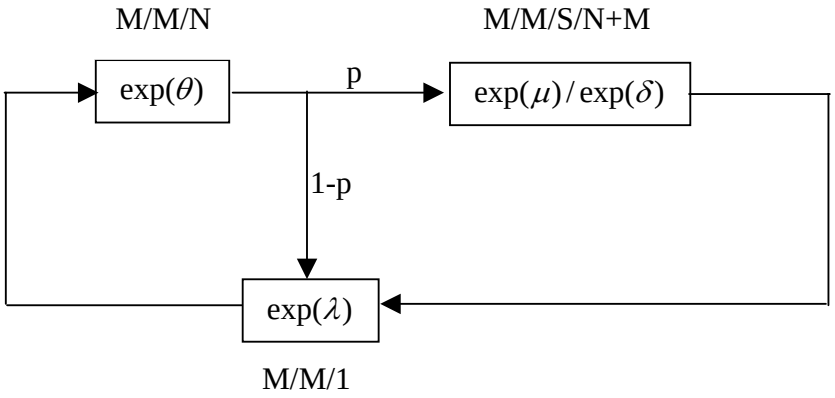


Figure 3.4: Schematic model of a call center with an interactive voice response, S agents and N trunk lines.

By normalization, we deduce the stationary probabilities $\pi(i,j)$ of having i calls at the IVR and j calls at the agents' station, which can be written in a normalized product form as follows:

$$\pi(i,j) = \begin{cases} \pi_0 \dfrac{1}{i!}\left(\dfrac{\lambda}{\theta}\right)^i \dfrac{1}{j!}\left(\dfrac{\lambda p}{\mu}\right)^j, & j \le S,\ \ 0 \le i+j \le N; \\ \pi_0 \dfrac{1}{i!}\left(\dfrac{\lambda}{\theta}\right)^i \dfrac{1}{S!}\left(\dfrac{\lambda p}{\mu}\right)^S \displaystyle\prod_{k=1}^{j-S} \dfrac{\lambda p}{S\mu + k\delta} & j > S,\ \ 0 \le i+j \le N; \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

where

$$\pi_0 = \left(\sum_{i=0}^{N-S-1} \sum_{j=S+1}^{N-i} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!}\left(\frac{\lambda p}{\mu}\right)^S \prod_{k=1}^{j-S} \frac{\lambda p}{S\mu + k\delta} + \sum_{i+j \le N, j \le S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{\lambda p}{\mu}\right)^j \right)^{-1}. \tag{3.8}$$

Formally, for all states (i,j), we have

$$\pi(i,j) = \lim_{t\to\infty} P\{Q_1(t) = i,\, Q_2(t) = j\}.$$

We say that the system is in state (k, j), $0 \leq j \leq k \leq N$, when it contains exactly k calls, and j is the number of calls in the agents' station (waiting or served); hence, $k - j$ is the number of calls in the IVR. The distribution function of the waiting time and the probability that a call starts its service immediately after leaving the IVR were found by Srinivasan et al. [80] and given by:

$$P(W \leq t) \triangleq 1 - \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k, j) \sum_{l=0}^{j-S} \frac{(\mu S t)^l e^{-\mu S t}}{l!} \tag{3.9}$$

and

$$P(W = 0) \triangleq \sum_{k=1}^{N} \sum_{j=0}^{min(k,S)-1} \chi(k, j) \tag{3.10}$$

where $\chi(k, j)$, $0 \leq j < k \leq N$, is the probability that the system is in state (k, j), given that a call (among the $k - j$ customers) is about to finish its IVR service:

$$\chi(k, j) = \frac{(k - j)\, \pi(k - j, j)}{\sum_{l=0}^{N} \sum_{m=0}^{l} (l - m)\, \pi(l - m, m)}. \tag{3.11}$$

The *probability of abandonment*, given waiting can be presented as follows

$$P(Ab|W > 0) = \frac{\sum_{j=S+1}^{N} \sum_{i=0}^{N-j} \pi(i, j)(j - S)\delta}{\sum_{j=S+1}^{N} \sum_{i=0}^{N-j} \pi(i, j)\, (S\mu + (j - S)\delta)}. \tag{3.12}$$

The conditional *expected waiting time* $E[W|W > 0]$ can be derived from (3.12) using the following property

$$E[W|W > 0] = \frac{P(Ab|W > 0)}{\delta}. \tag{3.13}$$

This relationship is well known for the M/M/S/N+M queue and one can easily show that it holds for the model with an IVR as well.

The fraction of the customers that wait in queue, which we refer to as the *delay probability*, is given by

$$P(W > 0) = \sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \chi(i, j). \tag{3.14}$$

Equation (3.14) gives the conditional probability that a calling customer does not immediately reach an agent, given that the calling customer is not blocked, i.e., $P(W > 0)$ is the *delay probability for served customers*. This conditional probability can be reduced to an unconditional probability via the "Arrival Theorem" [18]. Specifically, for the system with N trunk lines and S agents, the fraction of customers that are required to wait after their IVR service, coincides with the probability that a system with $N-1$ trunk lines and S agents has all its agents busy, namely

$$P_N(W > 0) = P_{N-1}(Q_2(\infty) \geq S). \tag{3.15}$$

3.2 Asymptotic Analysis in the QED Regime

3.2.1 The Domain for Asymptotic Analysis

All the following approximations will be derived when the arrival rate λ tends to infinity. In order for the system to not be overloaded, we assume that the number of agents S and the number of trunk lines N tend to infinity as well.

Our approximations for performance measures calculated according to the M/M/S/N+M queue model are the same as were formulated in [65]:

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S}{\sqrt{S}} = \eta, \quad \eta \geq 0, \\ &(ii) \quad \lim_{\lambda\to\infty} \sqrt{S}\left(1 - \frac{\lambda}{\mu S}\right) = \beta, \quad -\infty < \beta < \infty. \end{aligned} \tag{3.16}$$

The asymptotic domain for the model with an IVR were presented first in [50] and has the following form:

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty; \\ &(ii) \quad \lim_{\lambda\to\infty} \sqrt{S}\left(1 - \frac{\lambda p}{\mu S}\right) = \beta, \quad -\infty < \beta < \infty. \end{aligned} \tag{3.17}$$

3.2.2 The M/M/S/N+M Queue

We start with approximations for performance characteristics of the M/M/S/N+M queue. The results are formalized in the following theorem.

Theorem 3.2.1. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy conditions (3.16) where μ and δ are fixed.* [1] *Then the asymptotic behavior of the system is described in terms of the following performance measures:*

- *the asymptotic probability $P(W > 0)$ that a customer waits after the IVR process:*

$$\lim_{\lambda\to\infty} P(W>0) = \left(1 + \frac{\sqrt{\frac{\mu}{\delta}}\Phi(\beta)\varphi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)}{\varphi(\beta)\left[\Phi\Big(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\Big)-\Phi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)\right]}\right)^{-1}, \tag{3.18}$$

- *the asymptotic probability of abandonment, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(Ab|W>0) = \frac{\sqrt{\frac{\delta}{\mu}}\varphi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)}{\Phi\Big(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\Big)-\Phi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)} - \beta, \tag{3.19}$$

- *the asymptotic expectation of the waiting time, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}E[W|W>0] = \frac{1}{\delta}\frac{\sqrt{\frac{\delta}{\mu}}\varphi(\beta\sqrt{\frac{\mu}{\delta}})}{\Phi\Big(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\Big)-\Phi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)} - \frac{\beta}{\delta}, \tag{3.20}$$

- *the asymptotic probability of blocking:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(block) = \frac{\frac{\varphi(\beta)}{\varphi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)}\varphi\Big(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\Big)}{\Phi(\beta)+\sqrt{\frac{\delta}{\mu}}\frac{\varphi(\beta)}{\varphi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)}\left[\Phi\Big(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\Big)-\Phi\Big(\beta\sqrt{\frac{\mu}{\delta}}\Big)\right]}, \tag{3.21}$$

  where Φ and φ are the standard normal cumulative distribution and density functions, respectively.

[1] When $\eta = 0$, the M/M/S/N+M queue is equivalent to the M/M/S/S loss system. In this case $P(Ab|W > 0)$ and $E[W]$ are equal to 0 and their approximations are not relevant.

The proof of Theorem 3.2.1 is presented in Section 3.6.1 and it is carried out by using formulas (3.3) - (3.6), where the stationary probabilities are defined by (3.1) and (3.2).

3.2.3 Call Center with an IVR

In the following theorem we formulate approximations of the operational performance measures for a call center with an IVR, which were defined previously in Section 3.1.2. Proof of Theorem 3.2.2 is presented in Section 3.6.2.

Theorem 3.2.2. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED conditions* (3.17)*, where μ, p, θ and δ are fixed. Then the asymptotic behavior of the system is described in terms of the following performance measures:*

- *the asymptotic probability $P(W > 0)$ that a customer waits after the IVR process:*

$$\lim_{\lambda\to\infty} P(W > 0) = \left(1 + \frac{\gamma}{\xi_1 - \xi_2}\right)^{-1}, \tag{3.22}$$

- *the asymptotic probability of abandonment, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(Ab|W > 0) = \frac{\sqrt{\frac{\mu}{\delta}}\varphi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta)}{\int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta\sqrt{\frac{\mu}{\delta}} - t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt} - \beta, \tag{3.23}$$

- *the asymptotic expectation of waiting time, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}E[W|W > 0] = \frac{1}{\delta}\frac{\sqrt{\frac{\mu}{\delta}}\varphi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta)}{\int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta\sqrt{\frac{\mu}{\delta}} - t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt} - \frac{\beta}{\delta}, \tag{3.24}$$

- *the asymptotic probability of a busy signal:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(block) = \frac{\nu + \xi_2\varphi\Big[\Big(\eta + \beta\sqrt{\frac{p\mu\theta}{\delta}}\Big)/\Big(\sqrt{1 + \frac{p\theta}{\delta}}\Big)\Big]/\Big[1 - \Phi(\beta\sqrt{\frac{\mu}{\delta}})\Big]}{\gamma + \xi_1 - \xi_2}; \tag{3.25}$$

in the above,

$$\xi_1 = \sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})}\int_{-\infty}^{\eta}\Phi\left((\eta-t)\sqrt{\frac{\delta}{p\theta}}+\beta\sqrt{\frac{\mu}{\delta}}\right)\varphi(t)dt,$$

$$\xi_2 = \sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})}\Phi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta),$$

$$\gamma = \int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt, \text{ and } \nu = \frac{1}{\sqrt{1+\sqrt{\frac{\mu}{p\theta}}}}\varphi\left(\frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}\right)\Phi\left(\frac{\beta\sqrt{\frac{\mu}{p\theta}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}\right).$$

3.3 Accuracy of the Approximations

3.3.1 Approximations for the M/M/S/N+M Queue

Examining the approximations for performance measures of the M/M/S/N+M queue, we model a mid-sized call center, in which the arrival rate λ is 100 customers per minute. The number of agents S is in the domain where the traffic intensity $\rho = \frac{\lambda p}{\mu S}$ is about 1 (namely, the number of agents is between 80 and 120).

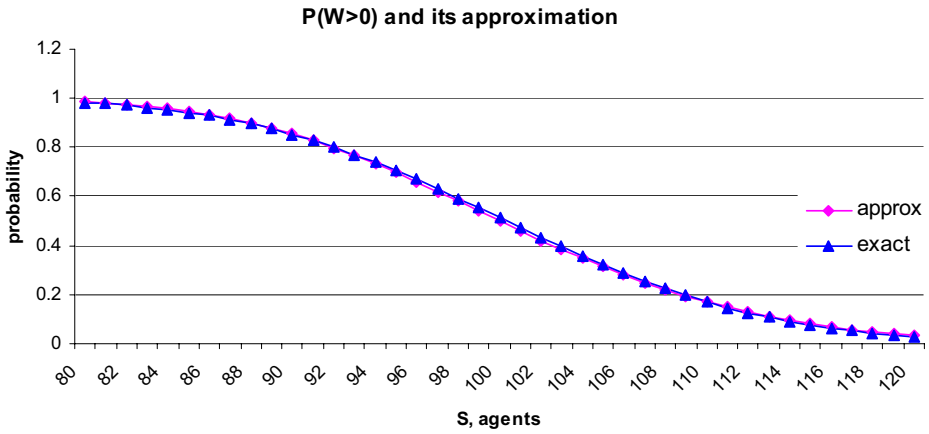


Figure 3.5: Comparison of the exact probability of waiting and its approximation, for a mid-sized call center with arrival rate 100 and 150 trunk lines.

We let $p = \mu = \theta = \delta = 1$. The number of trunk lines is mostly 150, but when we check the probability of blocking, we take the number of trunk lines to

be 120 (this in order to avoid very small values).

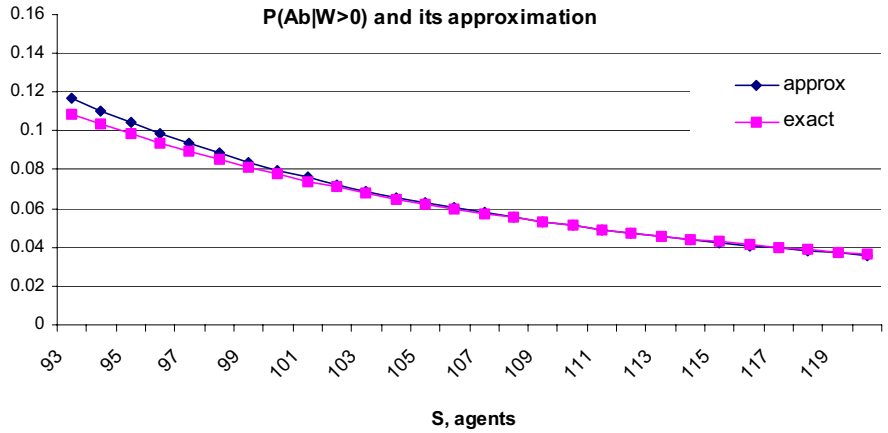


Figure 3.6: Comparison of the exact probability of abandonment, given waiting, and its approximation, for a mid-sized call center with arrival rate 100 and 150 trunk lines.

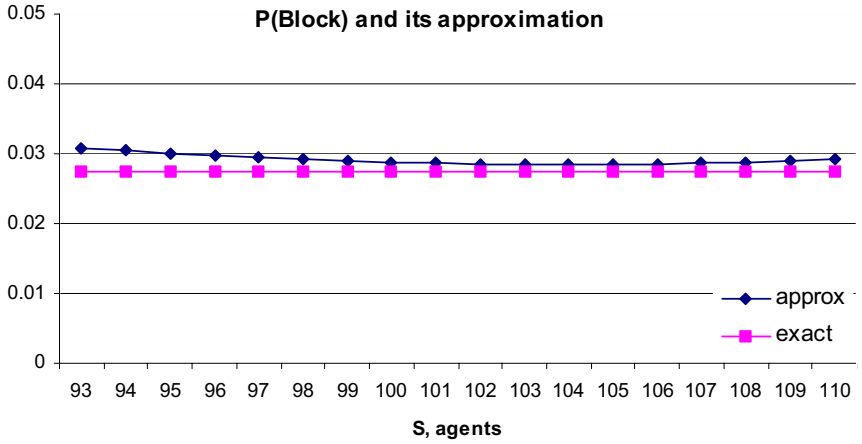


Figure 3.7: Comparison of the exact probability of finding the system busy and its approximation, for a mid-sized call center with arrival rate 100 and 120 trunk lines.

One of the conclusions which can be derived from Figures 3.5-3.7 is the fact that the approximations which were founded are close to the exact value although

in the small-sized call center. In addition, we have to emphasize that the calculation of the exact value is very difficult practically in the case of a bigger call center, for example, when the arrival rate λ is 500, the number of trunk lines N is 1500, and the number of agents S is between 450 and 550 agents. Using the fact that the approximation is very close to the exact value we can easily calculate the performance measures in such call centers.

3.3.2 Approximations of the Model with an IVR

The accuracy of approximations for a model *without* abandonment was provided in [50]. These approximations turn out to be extremely accurate, over a very wide range of parameters (S already from 10 and above, $N \geq 50$). Here, we present approximations that accommodate abandonments. The numerical analysis is heavier due to the increased number of integral-approximations. For example, the approximation of $P(W > 0)$ involves an integral in both γ and ξ (as opposed to only γ, in the model without abandonment). In addition, for calculations of the exact values we are restricted to relatively small N's ($N \leq 80$ here, as opposed to $N \leq 170$).

To investigate the performance of our approximations, we compare the performance measures of a model with an IVR and abandonment that corresponds to a small-sized call center that has the arrival rate λ of 30 customers per minute. The number of agents S is in the domain where the traffic intensity $\rho = \frac{\lambda p}{\mu S}$ is about 1 (namely, the number of agents is between 20 and 40, i.e. $S \approx 30 \pm 2 \cdot \sqrt{30}$). For simplicity, we let $p = \mu = \theta = \delta = 1$. The number of trunk lines is 80. For each value of the number of agents S, we calculate the parameters η and β by using (3.17).

Figures 3.8 and 3.9 depict the comparison of the exact probability of waiting and the conditional probability to abandon with their approximations. The approximations are clearly close to the exact values.

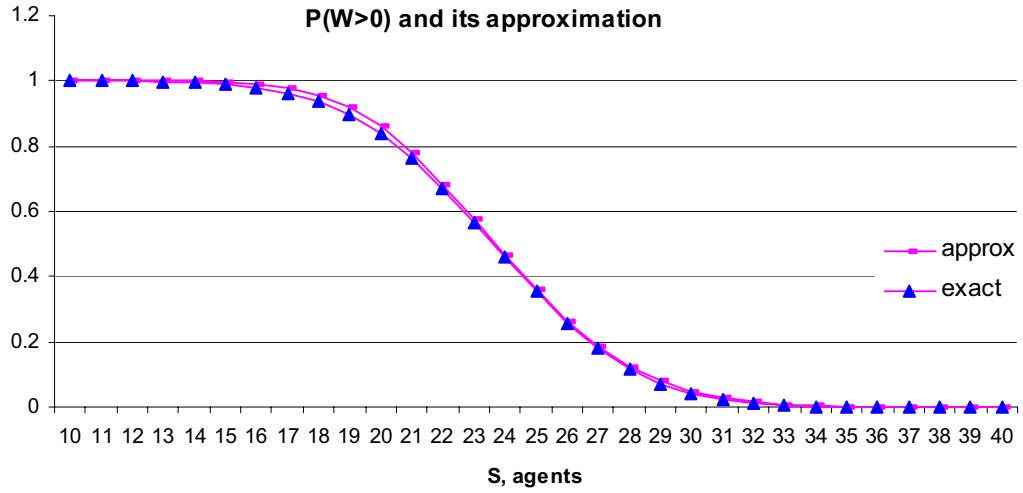


Figure 3.8: Comparison of the exact probability of waiting and its approximation (3.22) for a small-sized call center with arrival rate 30 and 80 trunk lines.

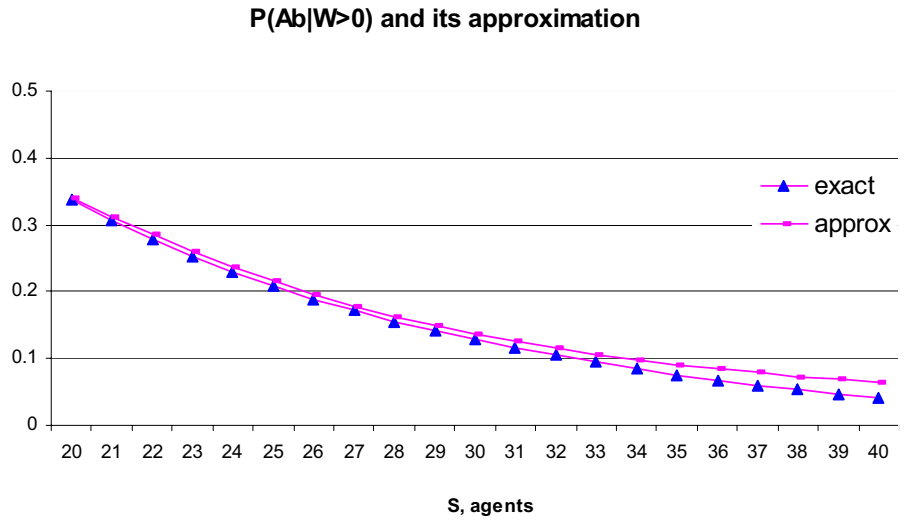


Figure 3.9: Comparison of the exact probability of abandonment, given waiting, and its approximation (3.23), for a small-sized call center with arrival rate 30 and 80 trunk lines.

Note, that

$$E[W|W>0] = \frac{1}{\delta} P(Ab|W>0).$$

Thus, it is expected that the approximation of $E[W]$ will also be close to the exact expectation.

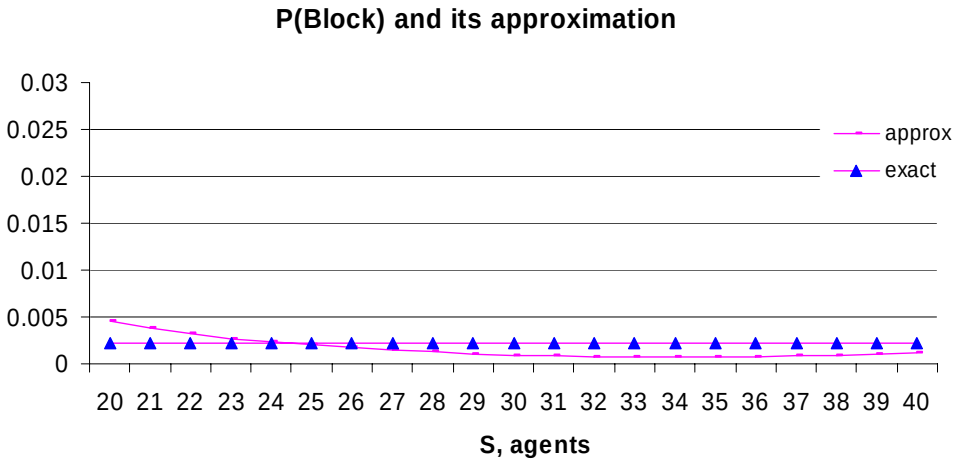


Figure 3.10: Comparison of the exact calculated probability to find all trunks busy and its approximation (3.25), for a mid-sized call center with arrival rate 30 and 80 trunk lines.

Figure 3.10 shows that the approximation of the probability of finding the system busy is accurate enough, and the differences are less than 0.002. One can thus argue that our approximation for the probability to find all trunks busy also works well.

3.4 Rules-of-Thumb

We derived approximations for performance measures in the QED regime (Quality and Efficiency Driven), as characterized by conditions (3.16) and (3.17). The detailed comparison in [50], between exact versus approximated performance, shows that the approximations often work perfectly, even *outside* the QED regime. In this section, we attempt to chart the boundary of this "outside" by summarizing our findings through practical rules-of-thumb (expressed via the offered load $R = \frac{\lambda}{\mu}$ for the M/M/S/N+M model or $R = \frac{\lambda p}{\mu}$ in the model with an IVR). These rules of thumb were derived via extensive numerical analysis of our analytical results.

3.4.1 Operational Regimes

As customary, one distinguishes three types of staffing regimes:

(ED) Efficiency-Driven, meaning under-staffing with respect to the offered load, to achieve high resource utilization;

(QD) Quality-Driven, meaning over-staffing with respect to the offered load, to achieve high service level;

(QED) Quality-and Efficiency-Driven, meaning rationalized staffing that carefully balances high levels of resource efficiency and service quality.

We shall use the characterization of the operational regimes, as formulated in [61] and presented in Table 3.1, in order to specify numerical ranges for the parameters β and η, in the $M/M/S/N$ queue and in the model with an IVR with and without abandonment. Specifying β corresponds to determining a staffing level, and specifying η corresponds to determining the number of trunk lines.

Table 3.1: Rules-of-thumb for operational regimes.

	ED	**QED**	**QD**
Staffing	$S \approx R - \gamma R$	$S \approx R + \beta\sqrt{R}$	$S \approx R + \gamma R$
% Delayed	$\approx 100\%$	constant over time (25%-75%)	$\approx 0\%$
% Abandoned	10% - 25%	1% - 5%	≈ 0
Average Wait	$\geq 10\% \cdot AST$	$\leq 10\% \cdot AST$	≈ 0

In Table 3.1, AST stands for Average Service Time.

3.4.2 System Parameters

The performance measures of a call center with an IVR, *without* abandonment, depends on β, η, $\frac{p\theta}{\mu}$ and S; in particular, large values of $\frac{p\theta}{\mu}$ and S improve performance (see [50] for elaboration). When one is adding abandonment to the system, one adds a parameter δ describing customers' patience. Large values of δ, corresponds to highly impatient customers, decrease the probability of waiting and the probability of blocking, but increase the probability of abandonment.

Small values of δ have the opposite influence. One must thus take into account 5 system's parameters. In order to reduce the dimension of this problem, we fix some parameters, at values that correspond to a realistic call center, based on our experience (see [83]):

IVR service time equals, on average, 1 minute;

Agents' service time equals, on average, 3 minutes;

Customers' patience, on average, takes values between 3 and 10 minutes;

Fraction of customers requesting agents' service, in addition to the IVR, equals 30%;

Offered load equals 200 Erlangs (200 minutes per minute).

Our goal is to identify the parameter values for η (determines the number of trunk lines) and β (determines the number of agents) that ensure QED performance as described in Table 3.1, while simultaneously estimating the value of the probability of blocking in each case (which does not appear in Table 3.1).

3.4.3 QED Regime in the M/M/S/N and M/M/S/N+M Queues

From the definition of the QED regime for the $M/M/S/N$ queue, η must be strictly positive ($\eta > 0$), because otherwise there would be hardly any queue and, thus, no reason to be concerned with the probability to wait or to abandon the system. Table 3.2 provides our rules-of-thumb for call centers without IVR and shows that when $\eta > 3$ the $M/M/S/N$ queue behaves as the $M/M/S$ queue (negligible blocking).

The rules-of-thumb presented in Table 3.2 were calculated under the assumption that the average customer patience equals 3 minutes (same as the average service time). As already noted, in practice this value can become much larger, but the performances are rather insensitive to the average patience time as long as the average $\leq$ 15 minutes. For higher average values the performances are similar to the corresponding model without abandonment.

Table 3.2: Rules-of-thumb for the QED regime in $M/M/S/N$ and $M/M/S/N+M$.

$S \approx R+\beta\sqrt{R}$ $N \approx S+\eta\sqrt{S}$	M/M/S/N	M/M/S/N+M
$0.5 \le \eta < 1.5$	$-1.5 < \beta < 0.5$	$-1.6 < \beta < 0.4$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.02, & \beta \ge 0; \end{cases}$	$\begin{cases} < -\beta/\sqrt{S}, & \beta < 0, \\ < 0.05, & \beta \ge 0; \end{cases}$
$1.5 \le \eta < 3$	$-0.5 < \beta < 0.8$	$-0.8 < \beta < 0.6$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.01, & \beta \ge 0; \end{cases}$	$\begin{cases} < 0.02, & \beta < 0 \\ \approx 0, & \beta \ge 0; \end{cases}$
$\eta > 3$	$\beta > 0$	$-0.5 < \beta < 0.8$
P(block)	≈ 0	≈ 0

3.4.4 QED Regime for a Call Center with an IVR with and without Abandonment

As in the previous subsection, the rules-of-thumb for the system with an IVR were calculated under the assumption that the average customer patience equals 3 minutes (same as the average service time). In the case where the system is with an IVR, there is no restrictions for η to be non negative, but we propose $\eta \ge 0$ because otherwise ($\eta < 0$), the probability of blocking is higher than 0.1. We believe that a call center cannot afford that 10% of its customers encounter a busy signal. Going the other way, a call center can extend the number of trunk lines to avoid the busy-line phenomenon altogether: as noted in Table 3.2, $\eta > 3$ suffices.

Table 3.3 shows that sometimes, one can reduce the number of trunk lines in order to improve service level. For instance, starting with $\eta > 3$ and the number of agents corresponding to $\beta = -0.8$ (ED performance), we can achieve QED performance by reducing the number of trunk lines via $\eta = 2$; in that way, we lose on waiting time and abandonment while the probability of blocking is still less than 0.01. Moreover, modern technology enables a message that replaces a busy-signal, with a suggestion to leave one's telephone number in order to be called back later; alternatively, a blocked call can be routed to an outsoursing alternative. Thus, we are not necessarily losing these "blocked" customers. See [52]

and [85] for an analysis where the asymptotically optimal number of trunk lines is determined.

Table 3.3: Rules-of-thumb for the QED regime in a call center with an IVR with and without abandonment.

$S \approx R + \beta\sqrt{R}$ $N \approx S + \frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}$	**IVR without abandonment**	**IVR with abandonment**
$0 \le \eta < 1$	$-1.2 < \beta < 0.2$	$-1.6 < \beta < 0$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.04, & \beta \ge 0; \end{cases}$	< 0.08
$1 \le \eta < 2$	$-0.7 < \beta < 0.5$	$-1.2 < \beta < 0.4$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.03, & \beta \ge 0; \end{cases}$	< 0.04
$2 \le \eta < 3$	$-0.3 < \beta < 0.7$	$-0.8 < \beta < 0.6$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.02, & \beta \ge 0; \end{cases}$	< 0.01
$\eta > 3$	$\beta > 0$	$-0.6 < \beta < 0.8$
P(block)	≈ 0	≈ 0

According to Table 3.3, when $\eta > 3$, the system with an IVR behaves as one with an infinite number of trunk lines.

3.4.5 QD and ED Regimes

For the QD and ED regimes (see Table 3.1), the number of agents can be specified via $0.1 \le \gamma \le 0.25$. In the case of QD, the number of agents is over-staffed; limiting the number of trunk lines will cause unreasonable levels of agents' idleness, hence $\eta \ge 3$ makes sense. In the case of ED, the number of agents is under-staffed, and we are interested in reducing the system's offered load. Therefore, we propose to take $\eta = 2$. This choice yields a probability of blocking to be approximately $\gamma/2$ (based on numerical experience).

3.4.6 Conclusions

Our rules-of-thumb demonstrate that for providing services in the QED regimes (in both cases: with and without an IVR) one requires the number of agents to be close to the system's offered load; the probability of blocking in the system with an IVR is always less than in the system without an IVR. One also observes that the existence of the abandonment phenomena considerably helps provide the same level of service as without abandonment, but with less agents. Moreover, as discussed in Section 3.4.4, it is possible to maintain operational service quality while reducing the number of agents by reducing access to the system. The cost is an increased busy signal. Hence, such a solution must result from a tradeoff between the probability of blocking and the probability to abandon.

3.5 Model Validation with Real Data

The approximations that have been developed can be of use in the operations management of a call center, for example when trying to maintain a pre-determined level of service quality. We analyze approximations of a real call center by models with and without an IVR. This evaluation is the goal of our empirical study, which is based on analyzing real data from a large call center. (The size of our call center, around 600-700 agents, forces one to use our approximations, as opposed to exact calculations which are numerically prohibitive.)

3.5.1 Data Description

The data for the current analysis come from a call center of a large U.S. bank - it will be referred to as the US Bank Call Center in the sequel. The full database archives all the calls handled by the call center over the period of 30 months from March 2001 until September 2003[2]. The call center consists of four different contact centers (nodes), which are connected using high technology switches so that, in effect, they can be considered as a single system. The call path can be described as follows. Customers, who make a call to the company, are first of all served in the IVR. After that, they either complete the call or choose to be served by an agent. In the latter case, customers typically listen to a message, after which they are routed, as will be now described, to one of the four call centers and join the agents' queue.

[2]The data is available at `http://seeserver.iem.technion.ac.il/see-terminal/`.

A schematic model of our US Bank Call Center is presented in Figure 3.11

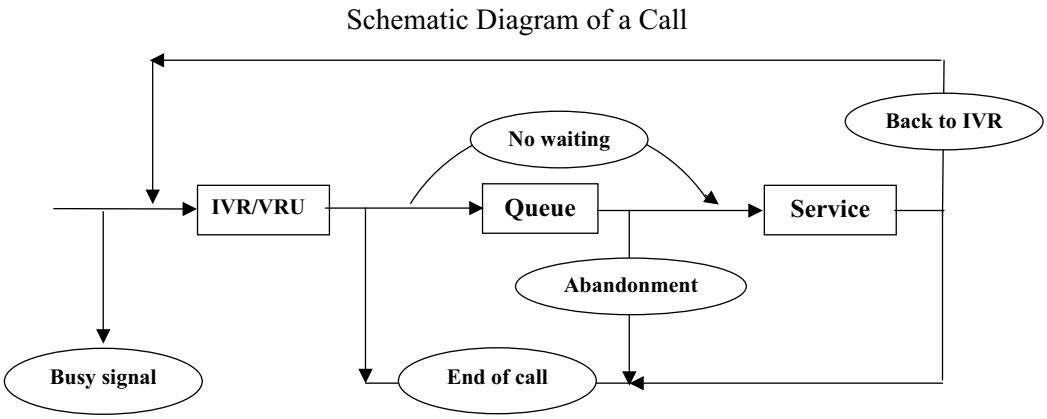


Figure 3.11: Schematic diagram of the call of a "Retail" customer in our US Bank call center.

The choice of routing is usually performed according to the customer's class, which is determined in the IVR. If all the agents are busy, the customer waits in the queue; otherwise, s/he is served immediately. Customers may abandon the queue before receiving service. If they wait in the queue of a specific node (one of the four connected) for more than 10 seconds, the call is transferred to a common queue - so-called "inter queue". This means that now the customer will be answered by an agent with an appropriate skill from any of the four nodes. After service by an agent, customers may either leave the system or return to the IVR, from which point a new *sub-call* ensues. The call center is relatively large with about 600 agents per shift, and is staffed 7 days a week, 24 hours a day.

3.5.2 Fitting the Theoretical Model to a Real System

Figure 3.11 describes the flow of a call through our call center. It differs somewhat from the models described in Section 3.1. The main difference is that it is possible for the customer to return to the IVR after being served by an agent. This is less common for so-called *Retail* customers who, almost as a rule, complete the call either after receiving service in the IVR or immediately after being served by an agent. We therefore neglect those few calls that return to the IVR and compare the models from Section 3.1 with the real system.

Our theoretical model assumes exponentially distributed service times in the IVR as well as for the agents. However, for the real data, neither of these service times have the exponential distribution. Figures 3.12 and 3.13, produced using the SEEStat program [83], display the distribution of service time in the IVR and agents' service time, respectively.

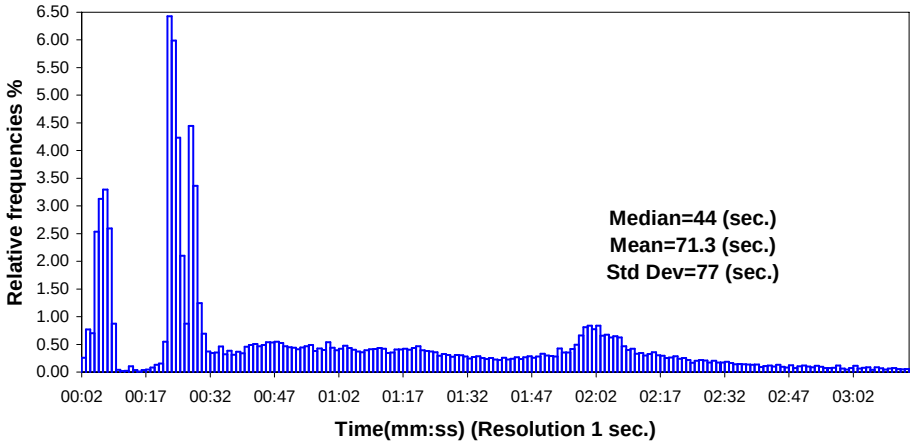


Figure 3.12: Histogram of the IVR service time for "Retail" customers

Figure 3.12 exhibits three peaks in the histogram of the IVR service time. The first peak can be attributed to calls of customers who are well familiar with the IVR menu and move fast to Agents' service; the second can be attributed to calls that, after an IVR announcement, opt for Agents' service; and the third peak can be related to the most common service in the IVR.

The distribution of the IVR service time is thus not exponential (see also [22]). A similar conclusion applies to agents' service time, as presented in Figure 3.13. Indeed, service time turns out to be log-normal (up to a probability mass near the origin) for about 93% of calls; the other 7% calls enjoy fast service for various reasons, for instance: mistaken calls, calls transferred to another service, unidentified calls sometimes transferred to an IVR, etc. (There are, incidentally, adverse reasons for short service times, for example agents "abandoning" their customers; see [16].)

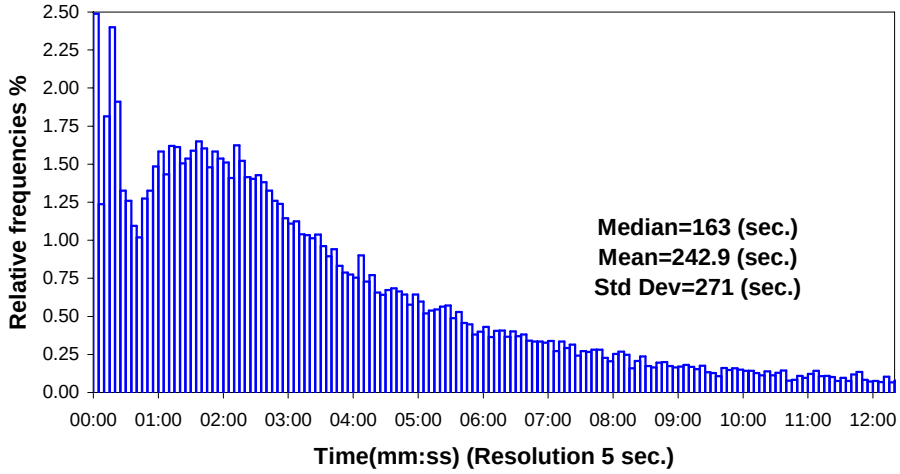


Figure 3.13: Histogram of Agents' service time for "Retail" customers

Similarly to non-Markovian (non-exponentially distributed) service times, the assumption that the arrival process is a homogeneous Poisson is also over simplistic. A more natural model for arrivals is an inhomogeneous Poisson process, as shown by Brown et al. [16], in fact modified to account for overdispersion (see [60]). However, and as done commonly in practice, if one divides the day into half-hour intervals, we get that within each interval the arrival rate is more or less constant and thus, within such intervals, we treat the arrivals as conforming to a Poisson process.

Even though most of the model assumptions do not prevail in practice, notably Markovian assumptions, experience has shown that Markovian models still provide *very useful* descriptions of non-Markovian systems (for example, the Erlang-A model in [16]). We thus proceed to validate our models against the US Bank Call Center, and our results will indeed demonstrate that this is a worthwhile insightful undertaking.

3.5.3 Comparison of Real and Approximated Performance Measures

For our calculations, the following variables must be estimated:

- λ - average arrival rate;
- θ - average rate of service in the IVR;
- μ - average rate of service by an agent;

- p - probability that a customer requests service by an agent;
- δ - average rate of customers' (im)patience;
- S - number of agents;
- N - number of trunk lines.

We consider the Retail service time distribution for April 12, 2001, which is an example of an ordinary week day. The analysis was carried out for data from calls arriving between 07:00 and 18:00. This choice was made since we were interested in investigating the system during periods of a meaningful load. We consider 30 minutes time intervals, since approximately 8000 calls are made during such intervals, we may expect that approximations for large λ would be appropriate. Moreover, system parameters seem to be reasonably constant over these intervals.

The following estimators will be calculated for each 30-minute interval as follows:

$$\widehat{\lambda} = \text{number of calls arriving to the system (30 min)}$$

$$\widehat{\theta} = \frac{30 \times 60}{\text{average IVR service time (sec)}}$$

$$\widehat{\mu} = \frac{30 \times 60}{\text{average agent service time (sec)}}$$

$$\widehat{p} = \frac{\text{number of calls seeking agent service}}{\widehat{\lambda}}$$

It should be noted that, strictly speaking, we are not calculating the actual average arrival rate because we see only the calls which did not find all trunks busy; practically, the fraction of customers that found all trunks busy is very small and hence the difference between the real and approximated (calculated our way) arrival rate is not significant.

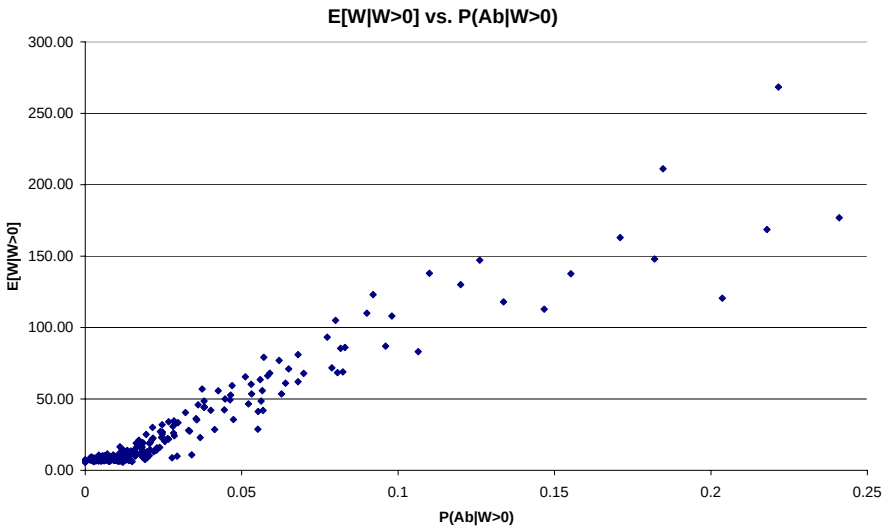


Figure 3.14: Relationship between the average waiting time given waiting, $E[W|W > 0]$, and the proportion of abandoned calls given waiting, $P(Ab|W > 0)$, for 30-minute intervals over 20 days.

The average rate of customers' patience was calculated via the relation

$$\delta = \frac{P(Ab|W > 0)}{E[W|W > 0]}, \tag{3.26}$$

which applies for the M/M/S/N+M queue (see [63] for details). Note that (3.26) assumes a linear relation between $P(Ab|W > 0)$ and $E[W|W > 0]$. Figure 3.14 demonstrates that this assumption is not unreasonable for our call center.

The estimation of the average rate of the customers' patience is thus the following:

$$\widehat{\delta} = \frac{\text{proportion of abandoned calls}}{\text{average of the waiting time (sec)}} \times 30 \times 60, \tag{3.27}$$

where both numerator and denominator are calculated for customers with a positive queueing time. Estimating the average rate of customers' patience for our data gave varying behavior of this parameter, for example at 14:30 its value is 5, at 15:00 it equals to 1, and at 15:30 it equals to 4. It is not unreasonable that customers' patience does not vary dramatically over each 30-minute period; hence, we smoothed the 30-minute values by using the R-function "smooth".

In order to use our approximations, we must assign an appropriate value for N, the number of trunk lines which is not available for us. We could consider the simplifying assumption that the number of trunk lines is unlimited. Certainly,

call centers are typically designed so that the probability of finding the system busy is very small, but nevertheless it is positive. One approach is to assume that, because the system is heavily loaded, there must be calls that are blocked since there are no explosions. In such circumstances, a naive way of underestimating N for each 30-minute period is as follows[3]:

$$\hat{N} = \frac{\text{total duration of all calls that arrived to the system}}{30 \cdot 60} .$$

The calculation of the number of agents is also problematic, because the agents who serve retail customers may also serve other types of customers, and vice versa: if all Retail agents are busy, the other agent types may serve Retail customers (see [57] for details). Thus, it is practically impossible to determine their exact number and that is why we use an averaged value, as follows:

$$\hat{S} = \frac{\text{total agent service time}}{30 \cdot 60}$$

Figure 3.15 compares the approximated theoretical probabilities of waiting based on the above estimators with the observed proportion of waiting customers, as estimated directly from the data. The dark blue curve (with diamonds) shows the proportion of customers that are waiting in the queue before agent service. This proportion is calculated for each half-hour period. The lilac curve (squares) shows the approximation based on the model with an IVR, calculated for each half-hour period. The blue curve (triangles) corresponds to the approximation based on the M/M/S/N+M queue model. We conclude that our approximations are performing reasonably well, especially based on the model with IVR. The approximate values for this model, in many intervals, are very close to the exact proportion. In some intervals the difference is about 10%, which can be attributed to the non-perfect correspondence between the model and the real call center. An additional explanation is in the estimation of the parameters, such as N and S, which we estimate in a very crude way. The approximation from the M/M/S/N+M queue works less well and sometimes it does not even reflect the trends seen for the real values: namely, where the real values decrease the approximation increases and vice versa. The reasons for these discrepancies can be the same as previously stated, as well as due to ignoring the IVR influence.

[3]Note that for the system with an IVR, $\hat{N}$ depends on the total duration of calls in the IVR, agents' queue and service. For the system without IVR, it depends only on the total duration of calls in the agents' queue and service.

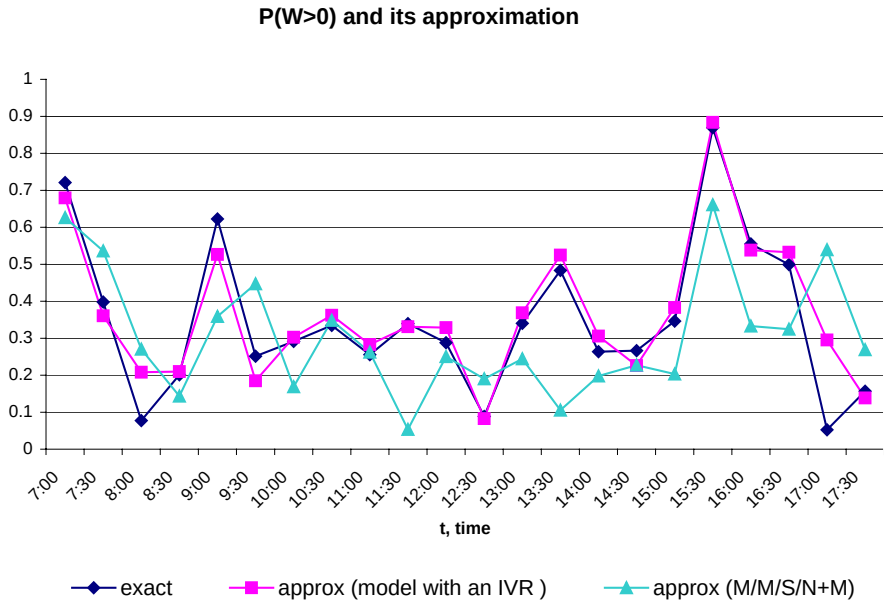


Figure 3.15: Comparison of approximate and observed probability of waiting.

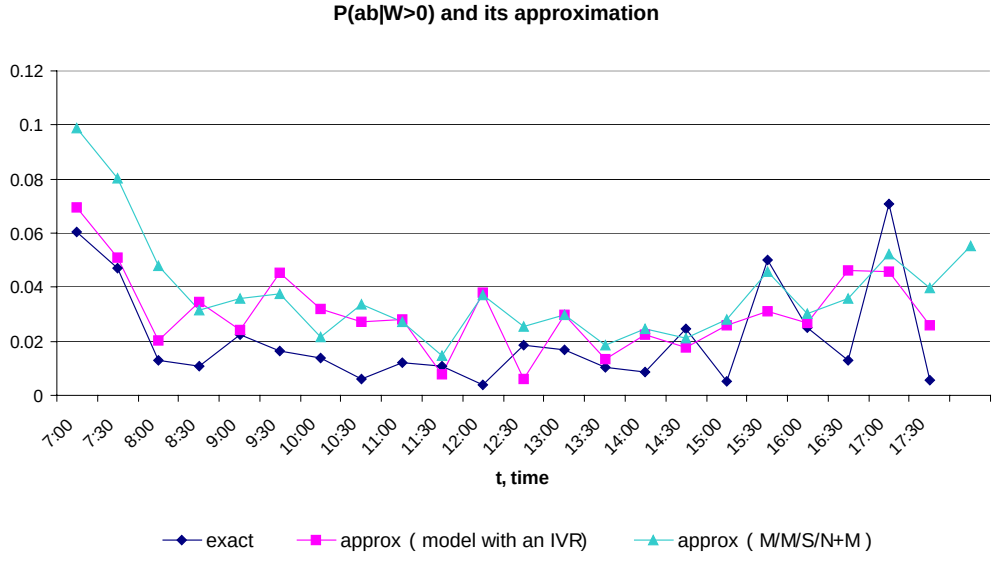


Figure 3.16: Comparison of the approximate and observed conditional probability to abandon $P(ab|W > 0)$.

In the Figures 3.16-3.17, we compare the observed and approximate condi-

tional probability for customer to abandon the system and the conditional average waiting time, given waiting.

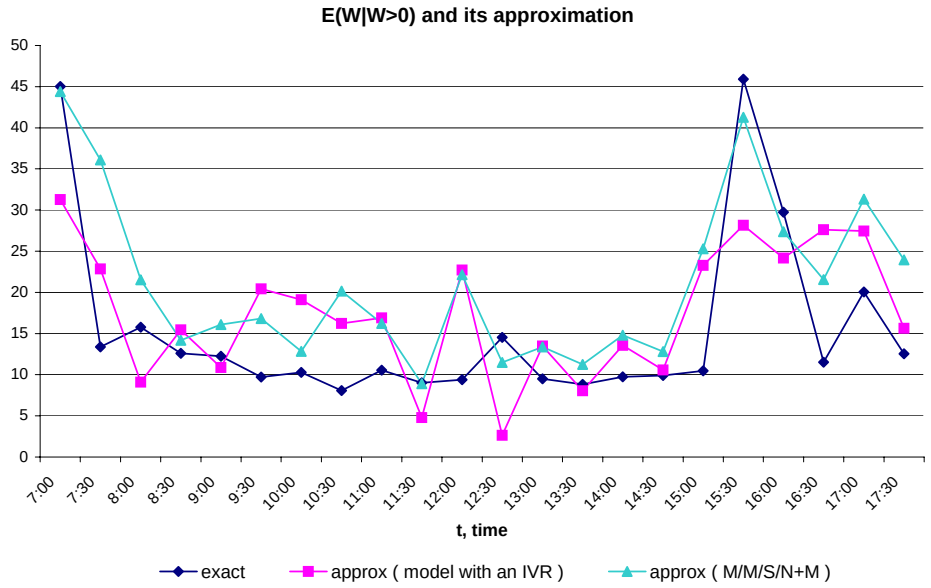


Figure 3.17: Comparison of the approximate and observed conditional average waiting time $E(W|W > 0)$, in seconds.

The conclusion based on Figures 3.16 and 3.17 are similar to those based on Figure 3.15. In some cases we see larger deviations, and a possible explanation is the sensitivity of our measures under heavy traffic, i.e. a little change of parameter values can dramatically change the performance measures.

In summary, both models considered above provide useful approximations to reality. Visual inspection reveals that the model with an IVR does it much better than the M/M/S/N+M queue.

3.6 Proofs

3.6.1 Proof of Theorem 3.2.1

Note, that when $i > S$ the probability $\pi(i)$ can be rewritten as follows

$$\pi(i) = \left(\frac{\lambda}{\mu}\right)^S \frac{1}{S!} \left(\frac{\lambda}{\delta}\right)^{i-S} \frac{\left(\frac{S\mu}{\delta}\right)!}{\left(\frac{S\mu}{\delta} + i - S\right)!} \pi(0). \tag{3.28}$$

Let us find the probability to wait. By using PASTA

$$P(W>0)=\sum_{i=S}^{N-1}\pi(i)=\frac{\displaystyle\sum_{i=S}^{N-1}\left(\frac{\lambda}{\mu}\right)^{S}\frac{1}{S!}\left(\frac{\lambda}{\delta}\right)^{i-S}\frac{\left(\frac{S\mu}{\delta}\right)!}{\left(\frac{S\mu}{\delta}+i-S\right)!}}{\displaystyle\sum_{i=0}^{S-1}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^{i}+\sum_{i=S}^{N}\left(\frac{\lambda}{\mu}\right)^{S}\frac{1}{S!}\left(\frac{\lambda}{\delta}\right)^{i-S}\frac{\left(\frac{S\mu}{\delta}\right)!}{\left(\frac{S\mu}{\delta}+i-S\right)!}}$$
$$=\frac{B_1(\lambda)}{A(\lambda)+B_2(\lambda)}.$$

Let us define $\xi_1(\lambda)=B_1(\lambda)e^{-\frac{\lambda}{\mu}}$, $\xi_2(\lambda)=B_2(\lambda)e^{-\frac{\lambda}{\mu}}$ and $\gamma(\lambda)=A(\lambda)e^{-\frac{\lambda}{\mu}}$. Then,

$$P(W>0)=\frac{\xi_1(\lambda)}{\gamma(\lambda)+\xi_2(\lambda)}.$$

Let us suppose that $S\mu/\delta$ is integer. This assumption is relaxed later.

$$\xi_1(\lambda)=\frac{\frac{1}{S!}\left(\frac{\lambda}{\mu}\right)^{S}e^{-\frac{\lambda}{\mu}}}{\frac{e^{-\frac{\lambda}{\delta}}}{\left(\frac{S\mu}{\delta}\right)!}\left(\frac{\lambda}{\delta}\right)^{\frac{S\mu}{\delta}}}\sum_{k=0}^{N-S-1}\frac{\left(\frac{\lambda}{\delta}\right)^{\frac{S\mu}{\delta}+k}e^{-\frac{\lambda}{\mu}}}{\left(\frac{S\mu}{\delta}+k\right)!}=\frac{P(Y=S)}{P(X=\frac{S\mu}{\delta})}P(X<\frac{S\mu}{\delta}+N-S),$$

where $X\overset{d}{=}Pois(\frac{\lambda}{\delta})$ and $Y\overset{d}{=}Pois(\frac{\lambda}{\mu})$. Using the Central Limit theorem, one can write that

$$P(X=\frac{S\mu}{\delta})\sim P\left(\frac{\frac{S\mu}{\delta}-\frac{\lambda}{\delta}}{\sqrt{\frac{\lambda}{\delta}}}-\frac{1}{\sqrt{\frac{\lambda}{\delta}}}<Y\leq\frac{\frac{S\mu}{\delta}-\frac{\lambda}{\delta}}{\sqrt{\frac{\lambda}{\delta}}}\right)$$
$$\sim\Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)-\Phi\left(\beta\sqrt{\frac{\mu}{\delta}}-\sqrt{\frac{\delta}{\lambda}}\right)\sim\sqrt{\frac{\delta}{\lambda}}\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right),$$

$$P(Y=S)\sim P\left(\frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}}-\sqrt{\frac{\lambda}{\mu}}<Y\leq\frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}}\right)$$
$$\sim\Phi\left(\beta\right)-\Phi\left(\beta-\sqrt{\frac{\mu}{\lambda}}\right)\sim\sqrt{\frac{\mu}{\lambda}}\varphi(\beta)$$

$$P\left(X\leq\frac{S\mu}{\delta}+N-S\right)\sim\Phi\left(\left(\frac{S\mu}{\delta}+\eta\sqrt{\frac{\lambda}{\mu}}-\frac{\lambda}{\delta}\right)/\sqrt{\frac{\lambda}{\delta}}\right)\sim\Phi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right).$$

Thus,

$$\lim_{\lambda\to\infty}\xi_1(\lambda)=\sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}\Phi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right).$$

Let us note that

$$\lim_{\lambda\to\infty}\xi_1(\lambda)=\lim_{\lambda\to\infty}\xi_2(\lambda),$$

and

$$\lim_{\lambda\to\infty}\gamma(\lambda)=\lim_{\lambda\to\infty}\sum_{i=0}^{S-1}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i e^{-\frac{\lambda}{\mu}}=\lim_{\lambda\to\infty}\Phi\left(\frac{S-1-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}}\right)=\Phi(\beta).$$

So,

$$\lim_{\lambda\to\infty}P(W>0)=\left(1+\frac{\Phi(\beta)\,\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}{\sqrt{\frac{\mu}{\delta}}\varphi(\beta)\Phi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)}\right)^{-1}. \tag{3.29}$$

Now, consider an approximation for $P(block)$. It can also be written as follows:

$$P(block)=\frac{\tau(\lambda)}{\gamma(\lambda)+\xi_2(\lambda)},$$

where

$$\tau(\lambda)=\frac{1}{S!}\left(\frac{\lambda}{\mu}\right)^S\left(\frac{\lambda}{\delta}\right)^{N-S}\frac{\left(\frac{S\mu}{\delta}\right)!}{\left(\frac{S\mu}{\delta}+N-S\right)!}=\frac{P(Y=S)}{P\left(X=\frac{S\mu}{\delta}\right)}P\left(X=\frac{S\mu}{\delta}+N-S\right).$$

Note, that

$$\begin{aligned}P\left(X=\frac{S\mu}{\delta}+N-S\right)&=P\left(\frac{\frac{S\mu}{\delta}+N-S-\frac{\lambda}{\delta}}{\sqrt{\frac{\lambda}{\delta}}}-\frac{1}{\sqrt{\frac{\lambda}{\delta}}}<X\le\frac{\frac{S\mu}{\delta}+N-S-\frac{\lambda}{\delta}}{\sqrt{\frac{\lambda}{\delta}}}\right)\\&=\frac{1}{\sqrt{\frac{\lambda}{\delta}}}\varphi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right).\end{aligned}$$

Then,

$$\begin{aligned}\lim_{\lambda\to\infty}\sqrt{S}\tau(\lambda)&=\lim_{\lambda\to\infty}\frac{1}{\sqrt{\frac{\lambda}{\mu}}}\varphi(\beta)\frac{\sqrt{\frac{\lambda}{\delta}}}{\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}\sqrt{\frac{\lambda}{\mu}}\varphi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)\\&=\frac{\varphi(\beta)\,\varphi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)}{\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}.\end{aligned}$$

Therefore,

$$\lim_{\lambda\to\infty}\sqrt{S}P(block)=\frac{\varphi\left(\beta\right)\varphi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)/\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}{\Phi\left(\beta\right)+\sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}\Phi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)}. \tag{3.30}$$

The conditional probability of abandonment can be written as follows

$$P(Ab|W>0)=1-\frac{\left(S\mu/\delta\right)\zeta_1(\lambda)}{\xi_2(\lambda)},$$

where $\xi_2(\lambda)$ as previously, and

$$\begin{aligned}\zeta_1(\lambda)&=\xi_2(\lambda)-P\left(X=\frac{S\mu}{\delta}\right)\\&\approx\frac{\delta}{\mu}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\mu/\delta})}\left[\Phi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)-\Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)-\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)\sqrt{\frac{\delta}{\lambda}}\right].\end{aligned}$$

Thus, we get

$$\lim_{\lambda\to\infty}\sqrt{S}P(Ab|W>0)=\frac{\sqrt{\frac{\mu}{\delta}}\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)-\beta\left[\Phi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)-\Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)\right]}{\Phi\left(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}\right)-\Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}.$$

3.6.2 Proof of Theorem 3.2.2

Approximation for $P(W>0)$.

According to (3.7), (3.8), (3.11) and (3.14), the operational characteristic $P(W>0)$ can be represented as follows:

$$P(W>0)=\left(1+\frac{A(\lambda)}{B(\lambda)}\right)^{-1}, \tag{3.31}$$

where

$$A(\lambda)=e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{i+j\le N-1,\ j\le S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j \tag{3.32}$$

and

$$B\left(\lambda\right)=e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{i=0}^{N-S-1}\sum_{j=S}^{N-i-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!}\left(\frac{p\lambda}{\mu}\right)^S\left(\frac{p\lambda}{\delta}\right)^{j-S}\frac{(S\mu/\delta)!}{(S\mu/\delta+j-S)!}. \tag{3.33}$$

We now derive QED approximations for $A(\lambda)$ and $B(\lambda)$, as λ, S and N tend to ∞, according to (3.17).

Approximation for $A(\lambda)$:

Consider a partition $\{S_j\}_{j=0}^{l}$ of the interval $[0, S]$:

$$S_j = S - j\Delta, \qquad j = 0, 1, ..., l; \quad S_{l+1} = 0, \tag{3.34}$$

where $\Delta = [\varepsilon\sqrt{\frac{\lambda p}{\mu}}]$, ε is an arbitrary non negative real and l is a positive integer.

If λ and S tend to infinity and satisfy the assumption (3.17)*(ii)*, then l is less than S/Δ for λ large enough and all the S_j belong to $[0, S]$, $j = 0, 1, ..., l$.

We emphasize that the length Δ of every interval $[S_{j-1}, S_j]$ depends on λ. The variable $A(\lambda)$ is given by formula (3.32), where the summation is taken over the trapezoid: $\{(i, j) \mid i \in [0, N - j] \ and \ j \in [0, S - 1]\}$, presented in Figure 3.18. Consider the lower estimate for $A(\lambda)$, given by the following sum:

$$\begin{aligned} A(\lambda) \geq A_1(\lambda) &= \sum_{k=0}^{l} \sum_{j=S_{k+1}}^{S_k - 1} \frac{1}{j!} \left(\frac{\lambda p}{\mu}\right)^j e^{-\frac{\lambda p}{\mu}} \sum_{i=0}^{N - S_k} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \\ &= \sum_{k=0}^{l} P(S_{k+1} \leq Z_\lambda < S_k) P(X_\lambda \leq N - S_k), \end{aligned} \tag{3.35}$$

where

$$Z_\lambda \stackrel{d}{=} Pois\left(\frac{\lambda p}{\mu}\right) \qquad \text{and} \qquad X_\lambda \stackrel{d}{=} Pois\left(\frac{\lambda}{\theta}\right). \tag{3.36}$$

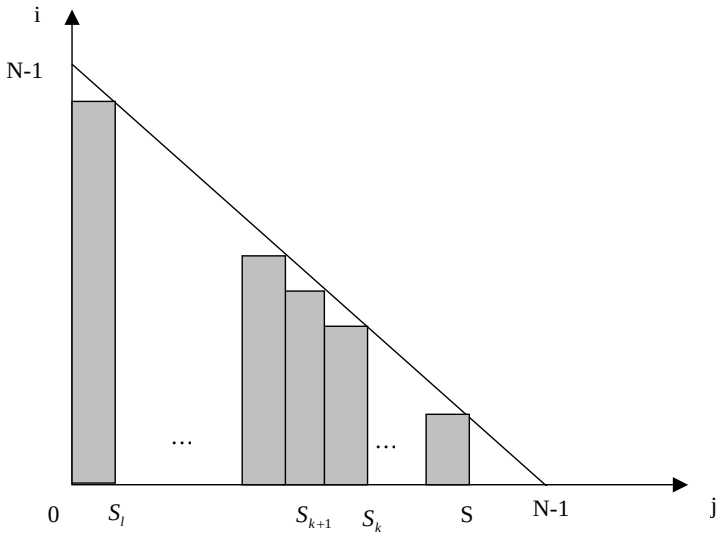


Figure 3.18: Area of the summation of the variable $A_1(\lambda)$.

Applying the Central Limit Theorem and making use of the relations

$$\lim_{\lambda\to\infty}\frac{S_k-\frac{\lambda p}{\mu}}{\sqrt{\frac{\lambda p}{\mu}}}=\beta-k\varepsilon,\quad \lim_{\lambda\to\infty}\frac{N-S_k-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}=\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}},\quad k=0,1,...,l, \tag{3.37}$$

one obtains

$$\lim_{\lambda\to\infty}P(S_{k+1}\le Z_\lambda<S_k)=\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon),\quad k=0,1,...l-1, \tag{3.38}$$

$$\lim_{\lambda\to\infty}P(0\le Z_\lambda<S_l)=\Phi(\beta-l\varepsilon), \tag{3.39}$$

$$\lim_{\lambda\to\infty}P(X_\lambda<N-S_k)=\Phi(\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}}),\quad k=0,1,...l. \tag{3.40}$$

It follows from (3.35) and (3.38), (3.39), (3.40) that

$$\begin{aligned}\liminf_{\lambda\to\infty}A(\lambda)\ge&\sum_{k=0}^{l-1}\Phi(\eta+k\varepsilon\sqrt{p\theta/\mu})[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]\\&+\Phi(\beta-l\varepsilon)\Phi(\eta+l\varepsilon\sqrt{p\theta/\mu}).\end{aligned} \tag{3.41}$$

It is easy to see that (3.41) is the lower Riemann-Stieltjes sum for the integral

$$-\int_0^\infty\Phi\left(\eta+s\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(\beta-s)=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt, \tag{3.42}$$

corresponding to the partition $\{\beta-k\varepsilon\}_{k=0}^{l}$ of the semi-axis $(-\infty,\beta)$.

Similarly, we obtain the upper Riemann-Stieltjes sum for the integral (3.42):

$$\limsup_{\lambda\to\infty}A(\lambda)\le\sum_{k=0}^{l-1}\Phi\left(\eta+(k+1)\varepsilon\sqrt{\frac{p\theta}{\mu}}\right)[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]+\Phi(\beta-l\varepsilon). \tag{3.43}$$

When $\varepsilon\to0$, the estimates (3.41), (3.43) lead to the following equality

$$\lim_{\lambda\to\infty}A(\lambda)=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt. \tag{3.44}$$

Approximation for $B(\lambda)$:

$$\begin{aligned}B(\lambda)&=\frac{\frac{e^{-\frac{p\lambda}{\mu}}}{S!}\left(\frac{p\lambda}{\mu}\right)^S}{\frac{e^{-\frac{p\lambda}{\delta}}}{(\frac{S\mu}{\delta})!}\left(\frac{p\lambda}{\delta}\right)^{\frac{S\mu}{\delta}}}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\sum_{k=0}^{N-S-i-1}\frac{\left(\frac{p\lambda}{\delta}\right)^{S\mu/\delta+k}e^{-\frac{p\lambda}{\delta}}}{(S\mu/\delta+k)!}\\&=\frac{P(Y_\lambda=S)}{P\left(X_\lambda=\frac{S\mu}{\delta}\right)}\sum_{i=0}^{N-S-1}P(Z_\lambda=i)\,P\left(\frac{S\mu}{\delta}\le X_\lambda\le N-S-i-1+\frac{S\mu}{\delta}\right),\end{aligned}$$

where

$$X_\lambda \stackrel{d}{=} Pois\left(\frac{p\lambda}{\delta}\right), \qquad Y_\lambda \stackrel{d}{=} Pois\left(\frac{p\lambda}{\mu}\right) \qquad \text{and} \qquad Z_\lambda \stackrel{d}{=} Pois\left(\frac{\lambda}{\theta}\right). \quad (3.45)$$

Analogously to calculations in approximation for $A(\lambda)$, with the use of the Central Limit Theorem we get

$$\begin{aligned}
&\lim_{\lambda\to\infty} \sum_{i=0}^{N-S-1} P\left(Z_\lambda = i\right) P\left(\frac{S\mu}{\delta} \le X_\lambda \le N-S-i-1+\frac{S\mu}{\delta}\right) \\
&\quad = \lim_{\lambda\to\infty} \sum_{i=0}^{N-S-1} P\left(Z_\lambda = i\right) \Big[\Phi\left(\frac{\frac{S\mu}{\delta}+N-S-i-1-\frac{p\lambda}{\delta}}{\sqrt{p\lambda/\delta}}\right) - \Phi\left(\frac{\frac{S\mu}{\delta}-\frac{p\lambda}{\delta}}{\sqrt{\frac{p\lambda}{\delta}}}\right)\Big] \\
&\quad = \lim_{\lambda\to\infty} \sum_{i=0}^{N-S-1} \Big[P\left(Z_\lambda \le i\right) - P\left(Z_\lambda \le i-1\right)\Big] \Phi\left((N-S-i)\sqrt{\frac{\delta}{p\lambda}} + \beta\sqrt{\frac{\mu}{\delta}}\right) \\
&\quad - \lim_{\lambda\to\infty} \Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right) \sum_{i=0}^{N-S-1} P\left(Z_\lambda = i\right) \\
&\quad = \lim_{\lambda\to\infty} \sum_{i=0}^{N-S-1} \Big[\Phi\left(\frac{i-\lambda/\theta}{\sqrt{\lambda/\theta}}\right) - \Phi\left(\frac{i-\lambda/\theta-1}{\sqrt{\lambda/\theta}}\right)\Big] \Phi\left((N-S-i)\sqrt{\frac{\delta}{p\lambda}} + \beta\sqrt{\frac{\mu}{\delta}}\right) \\
&\quad - \lim_{\lambda\to\infty} \Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right) P\left(Z_\lambda < N-S\right).
\end{aligned}$$

From condition (i) of (3.17) one can see that

$$\lim_{\lambda\to\infty} \Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right) P\left(Z_\lambda < N-S\right) = \Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right) \Phi\left(\eta\right). \quad (3.46)$$

For the first term we get

$$\begin{aligned}
&\lim_{\lambda\to\infty} \sum_{l=0}^{N-S-1} \Big[\Phi\left(\frac{N-S-l-\lambda/\theta}{\sqrt{\lambda/\theta}}\right) - \Phi\left(\frac{N-S-\lambda/\theta-(l+1)}{\sqrt{\lambda/\theta}}\right)\Big] \Phi\left(l\sqrt{\frac{\delta}{p\lambda}} + \beta\sqrt{\frac{\mu}{\delta}}\right) \\
&\quad = \lim_{\lambda\to\infty} - \sum_{l=0}^{N-S-1} \Delta\Phi\left(\eta - \frac{l}{\sqrt{\lambda}}\sqrt{\theta}\right) \Phi\left(\frac{l}{\sqrt{\lambda}}\sqrt{\frac{\delta}{p}} + \beta\sqrt{\frac{\mu}{\delta}}\right) \\
&\quad = -\int_0^\infty \Phi\left(t\sqrt{\frac{\delta}{p}} + \beta\sqrt{\frac{\mu}{\delta}}\right) d\Phi\left(\eta - t\sqrt{\theta}\right) = \int_{-\infty}^{\eta} \Phi\left((\eta-s)\sqrt{\frac{\delta}{p\theta}} + \beta\sqrt{\frac{\mu}{\delta}}\right) d\Phi\left(s\right).
\end{aligned} \quad (3.47)$$

It is easy to see that

$$P\left(Y=S\right)\sim P\left(\frac{S-\frac{p\lambda}{\mu}}{\sqrt{\frac{p\lambda}{\mu}}}-\frac{1}{\sqrt{\frac{p\lambda}{\mu}}}<Y\leq\frac{S-\frac{p\lambda}{\mu}}{\sqrt{\frac{p\lambda}{\mu}}}\right)\sim\frac{1}{\sqrt{\frac{p\lambda}{\mu}}}\varphi(\beta) \tag{3.48}$$

and

$$P\left(X=\frac{S\mu}{\delta}\right)\sim P\left(\frac{\frac{S\mu}{\delta}-\frac{p\lambda}{\delta}}{\sqrt{\frac{p\lambda}{\delta}}}-\frac{1}{\sqrt{\frac{p\lambda}{\delta}}}<Y\leq\frac{\frac{S\mu}{\delta}-\frac{p\lambda}{\delta}}{\sqrt{\frac{p\lambda}{\delta}}}\right)\sim\frac{1}{\sqrt{\frac{p\lambda}{\delta}}}\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right) \tag{3.49}$$

Combining (3.46)-(3.49) we get

$$\lim_{\lambda\to\infty}B(\lambda)=\sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)}\Big[\int_{-\infty}^{\eta}\Phi\left((\eta-t)\sqrt{\frac{\delta}{p\theta}}+\beta\sqrt{\frac{\mu}{\delta}}\right)d\Phi\left(t\right)-\Phi\left(\beta\sqrt{\frac{\mu}{\delta}}\right)\Phi\left(\eta\right)\Big].$$

Approximation for $P(Ab|W>0)$.

Note that the probability of abandonment, given waiting can be presented as follows

$$P(Ab|W>0)=1-\frac{S\mu}{p\lambda}\frac{C(\lambda)}{D(\lambda)},$$

where

$$\begin{aligned}C(\lambda)&=e^{-\lambda\left(\frac{1}{\theta}+\frac{p}{\delta}\right)}\sum_{j=S+1}^{N}\sum_{i=0}^{N-j}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^{i}\left(\frac{p\lambda}{\delta}\right)^{j-S+S\mu/\delta}\frac{1}{(S\mu/\delta+j-S)!}\\&=e^{-\lambda\left(\frac{1}{\theta}+\frac{p}{\delta}\right)}\sum_{k=1}^{N-S}\sum_{i=0}^{N-S-k-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^{i}\left(\frac{p\lambda}{\delta}\right)^{k+S\mu/\delta}\frac{1}{(S\mu/\delta+k)!}\\&=\sum_{k=1}^{N-S}\left(\frac{p\lambda}{\delta}\right)^{k+S\mu/\delta}\frac{e^{-\frac{p\lambda}{\delta}}}{(S\mu/\delta+k)!}\sum_{i=0}^{N-S-k-1}\frac{e^{-\frac{\lambda}{\theta}}}{i!}\left(\frac{\lambda}{\theta}\right)^{i}\\&=\sum_{k=1}^{N-S}P\left(X=\frac{S\mu}{\delta}+k\right)P\left(Y\leq N-S-k\right)\end{aligned}$$

and

$$
\begin{aligned}
D(\lambda) &= \sum_{j=S+1}^{N} \sum_{i=0}^{N-j} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \left(\frac{p\lambda}{\delta}\right)^{j-S-1+S\mu/\delta} \frac{e^{-\lambda\left(\frac{1}{\theta}+\frac{p}{\delta}\right)}}{(S\mu/\delta+j-S-1)!} \\
&= \sum_{k=1}^{N-S} \sum_{i=0}^{N-S-k-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \left(\frac{p\lambda}{\delta}\right)^{k+S\mu/\delta} \frac{e^{-\lambda\left(\frac{1}{\theta}+\frac{p}{\delta}\right)}}{(S\mu/\delta+k)!} + \frac{e^{-\frac{p\lambda}{\delta}}}{(S\mu/\delta)!} \sum_{i=0}^{N-S} \frac{e^{-\frac{\lambda}{\theta}}}{i!} \left(\frac{\lambda}{\theta}\right)^i \\
&= \sum_{k=1}^{N-S-1} P\left(X = \frac{S\mu}{\delta}+k\right) P\left(Y \le N-S-k-1\right) \\
&+ P\left(X = \frac{S\mu}{\delta}\right) P\left(Y \le N-S-1\right).
\end{aligned}
$$

In both expressions for $C(\lambda)$ and $D(\lambda)$ we assume that $X \stackrel{d}{=} Pois(\frac{p\lambda}{\delta})$ and $Y \stackrel{d}{=} Pois(\frac{\lambda}{\theta})$. Using conditions (3.1), the Central Limit Theorem, we obtain

$$
\begin{aligned}
\frac{N-S-k-\lambda/\theta}{\sqrt{\lambda/\theta}} &\sim \eta - \frac{k}{\sqrt{\lambda}}\sqrt{\theta}, \\
\frac{S\mu/\delta+k-p\lambda/\delta}{\sqrt{p\lambda/\delta}} &\sim \beta\sqrt{\frac{\mu}{\delta}} + \frac{k}{\sqrt{\lambda}}\sqrt{\frac{\delta}{p}}, \\
P\left(Y \le N-S-k\right) &\sim \Phi\left(\eta - k\sqrt{\frac{\theta}{\lambda}}\right)
\end{aligned}
$$

and

$$
P\left(X = \frac{S\mu}{\delta}+k\right) \sim \Phi\left(\beta\sqrt{\frac{\mu}{\delta}} - k\sqrt{\frac{\theta}{p\lambda}}\right) - \Phi\left(\beta\sqrt{\frac{\mu}{\delta}} - (k-1)\sqrt{\frac{\theta}{p\lambda}}\right).
$$

Then,

$$
\begin{aligned}
&\sum_{k=1}^{N-S} P\left(X = \frac{S\mu}{\delta}+k\right) P\left(Y \le N-S-k\right) \\
&\quad\sim \sum_{k=1}^{N-S-1} \Phi\left(\eta - k\varepsilon\sqrt{\theta}\right) \left[\Phi\left(\beta\sqrt{\frac{\mu}{\delta}} + k\varepsilon\sqrt{\frac{\delta}{p}}\right) - \Phi\left(\beta\sqrt{\frac{\mu}{\delta}} + (k-1)\varepsilon\sqrt{\frac{\delta}{p}}\right)\right] \\
&\quad\sim \int_0^\infty \Phi\left(\eta - t\sqrt{\theta}\right) d\Phi\left(\beta\sqrt{\frac{\mu}{\delta}} + t\sqrt{\frac{\delta}{p}}\right) \\
&\quad\sim \int_0^\infty \Phi\left(\eta - s\sqrt{\frac{p\theta}{\delta}}\right) d\Phi\left(\beta\sqrt{\frac{\mu}{\delta}} + s\right) \\
&\quad\sim -\beta \int_{\beta\sqrt{\frac{\mu}{\delta}}}^\infty \Phi\left(\eta - (t-\beta\sqrt{\frac{\mu}{\delta}})\sqrt{\frac{p\theta}{\delta}}\right) d\Phi\left(t\right).
\end{aligned}
$$

In addition, using

$$P\left(X = \frac{S\mu}{\delta}\right) \sim \frac{1}{\sqrt{p\lambda/\theta}}\varphi\left(\beta\sqrt{\frac{\mu}{\delta}}\right) \qquad \text{and} \qquad \frac{S\mu}{p\lambda} \sim \left(1 - \frac{\beta}{\sqrt{S}}\right)^{-1}$$

we get an approximation for the probability of abandonment given waiting, which is presented in Theorem 3.2.2.

3.7 Summary and Future Work

In this chapter we introduced appropriate models for the design of a typical call center. These models enable one to quantify the operational performance of a call center and to define staffing given a particular service level. Our proposed approximations of performance measures demonstrate a high level of accuracy and can be easily implemented for moderate-to-large call centers, where exact calculations are unsuccessful due to numerical instability.

The evaluation of approximations for the real call center data shows that even though most of the model assumptions do not prevail in practice, notably our Markovian assumptions, experience has shown that Markovian models still provide *very useful* descriptions of non-Markovian systems. The robustness of the M/M/S+M queue with respect to its characteristics were considered by Zeltyn [89] in his Ph.D. thesis. In particular, Zeltyn found that the relationships between the probability to abandon the system and the expectation of waiting time in M/M/S+G is the following: $P(Ab)/E[W] = f(0)$, where $f(0)$ is a value of customer patience density function at time 0. In the case of the M/M/S+M queue $f(0) = \theta$ and this is exactly the relationship that we used in our calculations.

In the future, we should like to improve our call center model by adding retrials, where by *retrials* we understand the customer's repeated attempts to receive the desired service after the initial failure to obtain it. This will make our model more realistic. Such an analysis can be extremely important, because the negative impact of customer retrials is the increase in system load and, hence, the deterioration of system performance and the corresponding increase in expenses. In whole, we should note that analysis of abandonment and retrial processes is very important for the management of call centers, because these phenomenons describe customers' satisfaction and the successfulness of the provided services.

One more realistic problem is the issue of different service requirements for different classes of customers. Such problems are called Skills-Based Routing and

they have already been investigated by Armony et al. [9] and Atar et al. [10]. It would be interesting to investigate the models of Skills-Based Routing for call centers with an IVR.

Chapter 4

CUSTOMER PATIENCE ANALYSIS

In many cases when a customer rings a call center s/he needs to wait in a queue before receiving someone to serve him/her. We can assume that each customer has a finite amount of time that s/he is ready to spend in the queue. If this time comes to an end and the customer has not been answered, s/he hangs up. In this chapter we provide an analysis of *customer patience*, which we define as his/her willingness to endure waiting in a queue before receiving service. The assessment of customer patience is a complicated issue because, in most cases, customers receive the required service before they lose their patience. The data with non-zero service time are called *censored data*, and these data require analysis of a special kind, known as survival analysis.

4.1 Description of the Data

We start with a short description of the data, which gives us the motivation for a model of customer patience considered in this chapter. The data we analyze are provided by a call center belonging to a financial company. From its call center, we have the data covering a period of almost three years, i.e. October 2006 - June 2009. The call center works twenty-four hours a day on weekdays (Sundays - Thursdays). It closes at 13:00 on Fridays, and reopens at about 17:00 on Saturdays. A customer making a call receives the service through an IVR or directly from an agent. After receiving the service provided by an IVR, the customer leaves the system or requests service from an agent. The customer requesting service from an agent is redirected to a pool of agents. If all the agents

are busy, the customer waits in a queue. Otherwise, s/he is served immediately. The customer is not always ready to wait in a queue, and he/she can choose to abandon the system at any point during the waiting period. After being served by an agent, the customer either finishes the call or proceeds to another service (another agent), and so on.

The call center in question provides various services. Some of them are similar in design and in their average service time. Others, on the contrary, are conceptually different. In our analysis, we will combine services of a similar kind into one group.

The data do not contain any personal information about customers, such as their age, social status, family status or education. Therefore, our analysis will be carried out only on the basis of the technical characteristics of the call. For each call, we have the following data:

- the individual number of a customer initiating this call (customer identification number),
- the type of customer (a type of priority given to the customer by the system),
- the beginning of each "call segment",
- the duration of each stage of a "call segment" (the service time, the waiting time in a queue or the post-call agent service time),
- the type of the service (an IVR service or an agent service that can include about fifteen different subtypes),
- the classification of call termination (after a received service, after call abandonment or due to a system error).

We identify each customer by his/her identification number which is retained in the field named "customer_id" and provides a unique number. However, sometimes "customer_id" can be unidentified or invalid. To avoid fake identification numbers we consider the data only for customers with fewer than thirty calls a month.

For the analysis of customer behavior, we use a notion of a "series" which we define as a sequence of consecutive calls from one customer happening in chronological order. If the time that elapses between two consecutive calls is less than three days we assume that these calls belong to the same "series", otherwise,

we assume that these calls belong to two different series. This separation is based on the assumption that a customer who has not called for a long time loses his/her experience with the system.

4.2 Model Selection

We propose a statistical model that can be used in the analysis of customer patience, under the setting of survival analysis. In our context, an *event* is the customer abandonment of the system before being served. For a customer who receives service, his/her patience time is not fully observed and is considered as *censored*. Hence, for each customer, at each call, the observed time is the time until abandonment (patience time) or time until being served, whichever comes first. The data to be used in the current research consists of customer calls with possibly multiple calls for a customer. We believe that the observed times of the same customer are not independent. Therefore, the Cox proportional hazard model [21] cannot be used directly, and we use a well-known and popular approach that deals with clustered data - the frailty model approach (Hougaard [42], Duchateau and Janssen [25], Aalen [1]).

The shared frailty model takes into account observed and unobserved personal factors of a customer. However, it is also reasonable to assume that the customer calls history influences his/her current waiting behavior. One of the models dedicated to such an analysis is the well studied recurrent events model and its extension to the shared frailty model [42]. However, these types of models cannot be applied directly in our case, since for typical recurrent event data, a subject can be censored at most once, and no information is available after this censoring time. In our data, a customer can call more than one time, and the response time in each call can be censored. So, we consider an extended shared frailty model assuming that customer patience changes with the number of the call, consistently for all customers.

4.3 Notation and Formulation of the Model

We consider n customers, where customer i has m_i calls in a series ($m_i \leq m$ for all $i = 1, ..., n$). Later, we consider a real data set analysis with a maximum of 5 calls for each customer ($m = 5$). We assume that the waiting behavior of each customer does not depend on the waiting behavior of other customers. Let T_{ij}^0

and C_{ij} denote the failure and censoring times, respectively, for call j of individual i $(i = 1, ..., n,\ j = 1, ..., m_i)$. The observed follow-up time is $T_{ij} = min\left(T_{ij}^0, C_{ij}\right)$, and the failure indicator is $\delta_{ij} = I\left(T_{ij}^0 \leq C_{ij}\right)$. For call j of customer i we observe a vector of covariates Z_{ij} and assume that the waiting behavior of customer i $(i = 1, ..., n)$ is influenced by some additional unobservable subject-dependent properties which are represented by the frailty variate w_i.

The conditional hazard function of the patience of customer i at the j-th call given the frailty w_i, is assumed to take the form

$$\lambda_{ij}(t) = \lambda_{0j}(t) w_i e^{\beta^T Z_{ij}} \qquad i = 1, ..., n \quad j = 1, ..., m_i, \tag{4.1}$$

where $\lambda_{0j}(t)$ is an unspecified baseline hazard function of call j and β is a p-dimensional vector of unknown regression coefficients. In this model, the baseline hazard functions are assumed to be different at each call, since it could be that customer behavior changes as he/she becomes more experienced with the system. It is also possible to consider a model with different regression coefficient vectors β_j, but for simplicity of presentation we suppose that $\beta_j = \beta$, for all j. We also assume the following assumptions:

(*a*) The frailty variate w_i is independent of m_i and Z_{ij} $\{j = 1, ..., m_i\}$.

(*b*) The frailty variates w_i $i = 1, ..., n$ are independent and identically distributed random variables with a density of known parametric form: $f(w) \equiv f(w; \theta)$, where θ is an unknown vector of parameters.

(*c*) The vector of covariates Z_{ij} is bounded.

(*d*) The random vectors $(m_i, T_{i1}^0, ..., T_{im_i}^0, C_{i1}, ..., C_{im_i}, Z_{i1}, ..., Z_{im_i}, w_i)$, $i = 1, ..., n$, are independent and identically distributed, and the model will be build conditional on m_i $i = 1, ..., n$.

(*e*) Given Z_{ij} $\{j = 1, ..., m_i\}$ and w_i, calls of customer i are independent.

(*f*) Given Z_{ij} $\{j = 1, ..., m_i\}$ and w_i, the censoring is independent and noninformative for w_i and (β, Λ_{0j}).

4.4 Estimation

The main goal of this work is to provide a test for comparing two or more baseline hazard functions. However, our proposed test requires estimators of the unknown

parameters: β, θ as well as $\{\lambda_{0j}(t)\}_{j=1}^{m}$. A simple estimation procedure that provides consistent estimators is given in the next section.

4.4.1 The Proposed Estimation Procedure

Our estimation procedure is based on the approach proposed by Gorfine et al. [37] which handles any frailty distribution with finite moments. We extend this method to the case of different baseline hazard functions $\lambda_{0j}(t)$. We describe in short the estimation procedure so that this work may be self-contained.

According to our model (4.1), the full likelihood can be written as

$$\begin{aligned} L &= \prod_{i=1}^{n} \int_0^{\infty} \prod_{j=1}^{m_i} \{\lambda_{ij}(T_{ij})\}^{\delta_{ij}} S_{ij}(T_{ij})\, f(w)dw \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m_i} \{\lambda_{0j}(T_{ij})\, e^{\beta^T Z_{ij}}\}^{\delta_{ij}} \prod_{i=1}^{n} \int_0^{\infty} w^{N_{i\cdot}(\tau)} e^{-wH_{i\cdot}(\tau)} f(w)dw, \end{aligned} \tag{4.2}$$

where τ is the end of the observation period, $N_{ij}(t) = \delta_{ij} I(T_{ij} \le t)$, $N_{i\cdot}(t) = \sum_{j=1}^{m_i} N_{ij}(t)$, $H_{ij}(t) = \Lambda_{0j}(T_{ij} \wedge t)\, e^{\beta^T Z_{ij}}$, $a \wedge b = min(a, b)$, $H_{i\cdot}(t) = \sum_{j=1}^{m_i} H_{ij}(t)$, $\Lambda_{0j}(t) = \int_0^t \lambda_{ij}(s)ds$ is the cumulative baseline hazard function and $S_{ij}(\cdot)$ is the conditional survival function for call j of subject i, namely,

$$S_{ij}(t) = \exp\left[-w_i e^{\beta^T Z_{ij}} \Lambda_{0j}(t)\right].$$

The log-likelihood is given by

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \delta_{ij} \ln\{\lambda_{0j}(T_{ij})\, e^{\beta^T Z_{ij}}\} + \sum_{i=1}^{n} \ln\Big\{ \int_0^{\infty} w^{N_{i\cdot}(\tau)} e^{-wH_{i\cdot}(\tau)} f(w)dw \Big\}. \tag{4.3}$$

As in [37], let $\gamma = \left(\beta^T, \theta\right)^T$, and for simplicity assume that θ is a scalar. If θ is a vector, the calculation can be derived in a similar way. The score vector, namely the vector of the log-likelihood derivatives with respect to γ, denoted by $U(\gamma, \left\{\Lambda_{0j}\right\}_{j=1}^{m}) = (U_1, ..., U_p, U_{p+1})$, is determined as follows

$$U_r = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[Z_{ijr} \big\{\delta_{ij} - H_{ij}(T_{ij})\big\} \frac{\int_0^{\infty} w^{N_{i\cdot}(\tau)+1} \exp\{-wH_{i\cdot}(\tau)\} f(w)dw}{\int_0^{\infty} w^{N_{i\cdot}(\tau)} \exp\{-wH_{i\cdot}(\tau)\} f(w)dw} \right]$$

for $r = 1, ..., p$, and

$$U_{p+1} = \frac{1}{n} \sum_{i=1}^{n} \frac{\int_0^{\infty} w^{N_{i\cdot}(\tau)} \exp\{-wH_{i\cdot}(\tau)\} f'(w)dw}{\int_0^{\infty} w^{N_{i\cdot}(\tau)} \exp\{-wH_{i\cdot}(\tau)\} f(w)dw},$$

where $f'(w) = df(w)/d\theta$. The estimation procedure consist of two steps. One is to estimate γ by substituting estimators of $\{\Lambda_{0j}\}_{j=1}^m$ into the equations

$$U(\gamma, \{\Lambda_{0j}\}_{j=1}^m) = 0.$$

The other is to estimate $\{\Lambda_{0j}\}_{j=1}^m$ given the estimated value of γ.

To this end, we provide here the estimators of $\{\Lambda_{0j}\}_{j=1}^m$. Define $Y_{ij}(t) = I(T_{ij} \geq t)$ and the entire observed history $\mathcal{F}_t$ up to time t as

$$\mathcal{F}_t = \sigma\{ N_{ij}(u),\ Y_{ij}(u),\ Z_{ij}, \quad i = 1, ..., n; \quad j = 1, ..., m_i; \quad 0 \leq u \leq t\}.$$

To simplify notation, we define $Z_{ij} = 0$ and $N_{ij}(t) = Y_{ij}(t) = 0$ for all $t \in [0, \tau]$ for each $m_i < j \leq m$ and $i = 1, ..., n$. As shown in Parner [70], applying the innovation theorem [14] to the observed history $\mathcal{F}_t$, the stochastic intensity process of $N_{ij}(t)$ with respect to $\mathcal{F}_t$ is given by

$$\lambda_{0j}(t) \exp(\beta^T Z_{ij}) Y_{ij}(t) \psi_i(t), \tag{4.4}$$

where

$$\psi_i(t) = E\Big(w_i \mid \mathcal{F}_{t-}\Big). \tag{4.5}$$

Using Bayes formula, we have

$$f(w_i \mid \mathcal{F}_{t-}) = \frac{w_i^{N_{i\cdot}(t-)} \exp\{-w_i H_{i\cdot}(t-)\} f(w_i)}{\int_0^\infty w_i^{N_{i\cdot}(t-)} \exp\{-w_i H_{i\cdot}(t-)\} f(w_i) dw_i}.$$

Therefore, the conditional expectation of w_i given the observed history at $[0, t)$ is as follows

$$\psi_i(t) = \frac{\int_0^\infty w^{N_{i\cdot}(t-)+1} e^{-w H_{i\cdot}(t-)} f(w) dw}{\int_0^\infty w^{N_{i\cdot}(t-)} e^{-w H_{i\cdot}(t-)} f(w) dw}. \tag{4.6}$$

It should be noted that $\psi_i(t)$ is a function of the unknown parameter γ and $\{\Lambda_{0j}\}_{j=1}^m$. Now, let

$$h_{ij}(t) = \psi_i(t) \exp(\beta^T Z_{ij}) \tag{4.7}$$

and note that given the intensity model (4.4), $h_{ij}(t)$ can be considered as a time-dependent covariate effect. Hence, the estimator of each Λ_{0j} is provided by using a Breslow-type [15] estimator as follows. Let $\hat{\Lambda}_{0j}$ be a step function with jumps at the observed failure times τ_{jk} ($k = 1, ..., K_j$ and $j = 1, ..., m$). Then, the jump size of $\hat{\Lambda}_{0j}$ at τ_{jk} given the value of $\hat{\gamma}$ is defined by

$$\Delta\hat{\Lambda}_{0j}(\tau_{jk}) = \frac{\sum_{i=1}^n dN_{ij}(\tau_{jk})}{\sum_{i=1}^n \hat{h}_{ij}(\tau_{jk}) Y_{ij}(\tau_{jk})}, \tag{4.8}$$

where $\hat{h}_{ij}(t) = \hat{\psi}_i(t)\exp(\hat{\beta}^T Z_{ij})$ and in $\hat{\psi}_i(t)$ we substitute $\hat{\gamma}$ and $\left\{\hat{\Lambda}_{0j}(t)\right\}_{j=1}^m$ into $\psi_i(t)$. It is important to note that each value $\Delta\hat{\Lambda}_{0j}(\tau_{jk})$ is a function of $\left\{\hat{\Lambda}_{0j}(t)\right\}_{j=1}^m$, where $t < \tau_{jk}$. Therefore, the estimation procedure is based on ordering the observed failure times of all the calls in increasing order and estimating $\left\{\Lambda_{0j}\right\}_{j=1}^m$ sequentially, according to the order of the observed failure times.

To summarize, the following is our proposed estimation procedure. Provide initial value of γ, and proceed as follows:

Step 1 : Given the value of γ estimate $\left\{\Lambda_{0j}\right\}_{j=1}^m$ by using (4.8).

Step 2 : Given the value of $\left\{\Lambda_{0j}\right\}_{j=1}^m$, estimate γ by solving

$$U(\gamma, \left\{\hat{\Lambda}_{0j}\right\}_{j=1}^m) = 0.$$

Step 3 : Repeat Steps 1 and 2 until convergence is reached with respect to $\left\{\hat{\Lambda}_{0j}\right\}_{j=1}^m$ and $\hat{\gamma}$.

For the choice of initial values for β we propose to use the naive Cox regression model, and for θ take 0. In case the integrals involved in (4.6) are not of closed analytical form, one can use numerical integration. As was already shown by Gorfine et al. [37], such an approach avoids the use of iterative processes in estimating the cumulative baseline hazard functions as required in other proposed procedures that are based on the EM-algorithm ([90], among others).

4.4.2 Asymptotic Properties

In this section, we formulate and summarize the asymptotic results of our proposed estimators. We denote by $\gamma^o = \left(\beta^{oT}, \theta^o\right)^T$ and $\mathbf{\Lambda_0^o}(t) = \left\{\Lambda_{0j}^o(t)\right\}_{j=1}^m$ the true values of β, θ and $\mathbf{\Lambda_0}(t) = \left\{\Lambda_{0j}(t)\right\}_{j=1}^m$, respectively.

Claim 4.4.1. *The estimator $\hat{\mathbf{\Lambda}}_\mathbf{0}(t)$ converges almost surely to a limit $\mathbf{\Lambda_0}(t,\gamma)$ uniformly in t and γ, with $\mathbf{\Lambda_0}(t,\gamma) = \mathbf{\Lambda_0^o}(t)$, and $n^{1/2}[\hat{\mathbf{\Lambda}}_\mathbf{0}(t) - \mathbf{\Lambda_0^o}(t)]$ converges weakly to a Gaussian process.*

Claim 4.4.2. *The function $U[\gamma, \hat{\mathbf{\Lambda}}_\mathbf{0}(\cdot)]$ converges almost surely in t and γ to a limit $u[\gamma, \mathbf{\Lambda_0}(\cdot)]$.*

Claim 4.4.3. *There exists a unique consistent root to $U[\hat{\gamma}, \hat{\mathbf{\Lambda}}_\mathbf{0}(\cdot)] = 0$.*

Claim 4.4.4. *The asymptotic distribution of* $n^{1/2}\left(\hat{\gamma}-\gamma^{o}\right)$ *is normal with mean zero and with a covariance matrix that can be consistently estimated by a sandwich estimator.*

The proofs of Claims 4.4.1 - 4.4.4 along with all the required conditions are almost identical to those presented in Gorfine et al. [37] and Zucker et al. [91], since the only minor difference is the use of $\{\hat{\Lambda}_{0j}(t)\}_{j=1}^{m}$ instead of a global estimator based on all the calls together. Hence, the proofs and a detailed list of the additional required assumptions are omitted.

It should be noted that although a consistent variance estimator of $\hat{\gamma}$ and $\left\{\hat{\Lambda}_{0j}(t)\right\}_{j=1}^{m}$ can be provided, its form is very complicated. Hence, we recommend using the bootstrap approach.

4.5 Family of Weighted Tests for Correlated Samples

4.5.1 Introduction and preliminaries

Our main objective is to provide a test statistic for comparing the cumulative baseline hazard functions corresponding to different calls. Namely, we are interested in testing the hypothesis

$$H_0: \qquad \Lambda_{01}=\Lambda_{02}=...=\Lambda_{0m}=\Lambda_0, \tag{4.9}$$

where Λ_0 is some unspecified cumulative hazard with $\Lambda_0(t)<\infty$. As noted earlier, the intensity processes of the counting processes $N_{ij}(t)$ $i=1,...,n$, $j=1,...,m_i$, with respect to $\mathcal{F}_t$ has the form

$$h_{ij}(t)Y_{ij}(t)\lambda_{0j}(t). \tag{4.10}$$

However, given the frailty variate w_i, the intensity processes of $N_{ij}(t)$ $i=1,...,n$, $j=1,...,m_i$ take the form

$$\tilde{h}_{ij}(t)Y_{ij}(t)\lambda_{0j}(t) \tag{4.11}$$

with

$$\tilde{h}_{ij}(t)=w_i\exp(\beta^T Z_{ij}). \tag{4.12}$$

Let $\bar{Y}_j(t,\gamma)=\sum_{i=1}^{n}h_{ij}(t)Y_{ij}(t)$ and $\tilde{Y}_j(t,\gamma)=\sum_{i=1}^{n}\tilde{h}_{ij}(t)Y_{ij}(t)$, and note that $E\Big[\sum_{i=1}^{n}w_iY_{ij}(t)e^{\beta^T Z_{ij}}\Big]=E\Big[\sum_{i=1}^{n}E\Big(w_i\mid\mathcal{F}_{t-}\Big)Y_{ij}(t)e^{\beta^T Z_{ij}}\Big]$. Then, by the uniform strong law of large numbers [7] the functions $n^{-1}\bar{Y}_j(t,\gamma)$ and $n^{-1}\tilde{Y}_j(t,\gamma)$ converge to the same function, if one of them converges.

For deriving the asymptotic properties of our proposed test statistic, we make the following assumptions:

1. $\hat{W}_n(s)$ is nonnegative, cadlag or caglad, with bounded total variation, and converges in probability to some uniformly bounded integrable function $W(s)$, that is $\sup_{s\in[0,\tau]} \mid \hat{W}_n(s) - W(s) \mid \to 0$.

2. There exist positive deterministic functions $\bar{y}_j(s)$, $j = 1, ..., m$, such that

$$\sup_{s\in[0,\tau]} \mid n^{-1}\bar{Y}_j(s,\gamma^o) - \bar{y}_j(s) \mid \to 0 \qquad \sup_{s\in[0,\tau]} \mid n^{-1}\tilde{Y}_j(s,\gamma^o) - \bar{y}_j(s) \mid \to 0,$$

   $j = 1, ..., m$ almost surely, as $n \to \infty$.

3. $Q_{lj}(s,\gamma^o) = \frac{\partial}{\partial\gamma_l}\Big[\bar{Y}_j(s,\gamma)/\bar{Y}_\cdot(s,\gamma)\Big]_{\gamma=\gamma^o}$ $l = 1, ..., p+1$ $j = 1, ..., m$ are bounded over $[0,\tau]$ where $\bar{Y}_\cdot(s,\gamma^o) = \sum_{j=1}^m \bar{Y}_j(s,\gamma^o)$.

4. There exist deterministic functions $g_{lj}(s)$, $l = 1, ..., p+1$ $j = 1, ..., m$, such that

$$\sup_{s\in[0,\tau]} \mid Q_{lj}(s,\gamma^o) - g_{lj}(s) \mid \to 0$$

   almost surely, as $n \to \infty$.

4.5.2 Test for Equality of Two Hazard Functions

We start by comparing the cumulative baseline hazard functions of two calls. In this subsection we use indexes 1 and 2 for comparing any two baseline hazard functions out of the m possible functions. The extension to more than two calls will follow. Assume we are interested in testing the hypothesis

$$H_0: \qquad \Lambda_{01} = \Lambda_{02} = \Lambda_0. \tag{4.13}$$

We propose to use the weighted log-rank statistic (Fleming and Harrington [28]) that takes the form

$$\begin{aligned} S_n(t,\hat{\gamma}) &= \frac{1}{\sqrt{n}}\int_0^t \hat{W}_n(s)\frac{\bar{Y}_1(s,\hat{\gamma})\bar{Y}_2(s,\hat{\gamma})}{\bar{Y}_1(s,\hat{\gamma})+\bar{Y}_2(s,\hat{\gamma})}\Big\{d\hat{\Lambda}_{01}(s) - d\hat{\Lambda}_{02}(s)\Big\} \\ &= \frac{1}{\sqrt{n}}\int_0^t \hat{W}_n(s)\frac{\bar{Y}_1(s,\hat{\gamma})\bar{Y}_2(s,\hat{\gamma})}{\bar{Y}_1(s,\hat{\gamma})+\bar{Y}_2(s,\hat{\gamma})}\Big\{\frac{d\bar{N}_1(s)}{\bar{Y}_1(s,\hat{\gamma})} - \frac{d\bar{N}_2(s)}{\bar{Y}_2(s,\hat{\gamma})}\Big\}, \end{aligned} \tag{4.14}$$

for $t \in [0,\tau]$ where $d\bar{N}_j(s) = \sum_{i=1}^n dN_{ij}(s)$ and the estimators $\hat{\gamma}$ and $\{\hat{\Lambda}_{0j}\}_{j=1}^m$ are given in Section 4.4.

Given w_i and the intensity process (4.11), the process

$$M_{ij}(t) = N_{ij}(t) - w_i \int_0^t \lambda_{0j}(u) e^{\beta^T Z_{ij}} Y_{ij}(u) du$$

is a mean-zero martingale with respect to $\mathcal{F}_t$, namely

$$\begin{aligned} E\left[dM_{ij}(t) \mid w_i,\ \mathcal{F}_{t-}\right] &= E\left[dN_{ij}(t) \mid w_i,\ \mathcal{F}_{t-}\right] \\ &\quad - E\left[\lambda_{0j}(t) e^{\beta^T Z_{ij}} Y_{ij}(t) w_i dt \mid w_i,\ \mathcal{F}_{t-}\right] = 0. \end{aligned}$$

Then, given $w_\cdot = \{w_i\}_{i=1}^n$, the sum of these martingales $\bar{M}_j(t) = \sum_{i=1}^n M_{ij}(t)$ is also a mean-zero martingale with respect to $\mathcal{F}_{t-}$. Since $N_{i1}(t)$ and $N_{i2}(t)$ are conditionally independent given w_i for all $i = 1, ..., n$, then, given $w_\cdot$, $\bar{M}_1(t)$ and $\bar{M}_2(t)$ are uncorrelated martingales.

To simplify notation we define $\bar{Y}_\cdot(s,\gamma) = \bar{Y}_1(s,\gamma) + \bar{Y}_2(s,\gamma)$ and

$$\mathcal{G}(s,\gamma) = \frac{\bar{Y}_1(s,\gamma)\bar{Y}_2(s,\gamma)}{\bar{Y}_\cdot(s,\gamma)},\ D_n(s,\gamma) = \frac{\hat{W}_n(s)}{\sqrt{n}}\mathcal{G}(s,\gamma),\ D_j^n(s,\gamma) = \frac{\hat{W}_n(s)}{\sqrt{n}}\frac{\mathcal{G}(s,\gamma)}{\bar{Y}_j(s,\gamma)}$$

for $j = 1, 2$. For the asymptotic distribution of our test statistic $S_n(t,\hat{\gamma})$ and its variance estimator, we start with the following theorem.

Theorem 4.5.1. *Given Assumptions 3-4 the test statistic $S_n(t,\hat{\gamma})$ presented in* (4.14) *has the same asymptotic distribution as*

$$\tilde{S}_n(t,\gamma^o) + S_n^{**}(t), \tag{4.15}$$

where

$$\tilde{S}_n(t,\gamma^o) = \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s)\mathcal{G}(s,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\}, \tag{4.16}$$

$$S_n^{**}(t) = \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s)\mathcal{G}(s,\hat{\gamma})\Big\{\frac{\tilde{Y}_1(s,\gamma^o)d\Lambda_{01}(s)}{\bar{Y}_1(s,\hat{\gamma})} - \frac{\tilde{Y}_2(s,\gamma^o)d\Lambda_{02}(s)}{\bar{Y}_2(s,\hat{\gamma})}\Big\}. \tag{4.17}$$

The proof of Theorem 4.5.1 is presented in Section 4.7.1.

Now, consider the random variable $S_n^{**}(t)$. By the first order Taylor expansion about γ^o we get

$$\begin{aligned} S_n^{**}(t) &= \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s)\Big[\frac{\bar{Y}_2(s,\hat{\gamma})\tilde{Y}_1(s,\gamma^o)}{\bar{Y}_\cdot(s,\hat{\gamma})}d\Lambda_{01}(s) - \frac{\bar{Y}_1(s,\hat{\gamma})\tilde{Y}_2(s,\gamma^o)}{\bar{Y}_\cdot(s,\hat{\gamma})}d\Lambda_{02}(s)\Big] \\ &\approx \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s)\Big[\frac{\bar{Y}_2(s,\gamma^o)\tilde{Y}_1(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}d\Lambda_{01}(s) - \frac{\bar{Y}_1(s,\gamma^o)\tilde{Y}_2(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}d\Lambda_{02}(s)\Big] \\ &\quad + \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s)\Big\{\tilde{Y}_1(s,\gamma^o)\mathbf{Q_2}^T(s,\gamma^o)d\Lambda_{01}(s) \\ &\quad - \tilde{Y}_2(s,\gamma^o)\mathbf{Q_1}^T(s,\gamma^o)d\Lambda_{02}(s)\Big\}(\hat{\gamma} - \gamma^o). \end{aligned} \tag{4.18}$$

where

$$\mathbf{Q}_j^T(s,\gamma^o) = (Q_{1j},...,Q_{(p+1)j})^T \quad \text{and} \quad Q_{lj} = \frac{\partial}{\partial \gamma_l}\Big[\frac{\bar{Y}_j(s,\gamma)}{\bar{Y}_\cdot(s,\gamma)}\Big]_{\gamma=\gamma^o}$$

for $l = 1,..,p+1$ and $j = 1,2$.

The second term of the right-hand side of (4.18) represents the additional variability of $S_n(t,\hat{\gamma})$ due to the estimation of $\hat{\gamma}$ and, based on Claim 4.4.4 it is easy to see that it is asymptotically normal with mean zero. However, this term is expected to be of a negligible contribution to the total variance, since, $\hat{\gamma}$ is being estimated parametrically (Acar et al. [5], Section 2.3). It should be noted that our extensive simulation study presented in Section 4.8 also supports this argument. To summarize, we formulate the following conclusion.

Conclusion 4.5.1. *An approximation of the asymptotic distribution of $S_n(t,\hat{\gamma})$ is the asymptotic distribution of*

$$\tilde{S}_n^*(t,\gamma^o) = \tilde{S}_n(t,\gamma^o) + S_n^*(t,\gamma^o), \tag{4.19}$$

where

$$S_n^*(t,\gamma^o) = \frac{1}{\sqrt{n}}\int_0^t \hat{W}_n(s)\mathcal{G}(s,\gamma^o)\Big[\frac{\tilde{Y}_1(s,\gamma^o)}{\bar{Y}_1(s,\gamma^o)}d\Lambda_{01}(s) - \frac{\tilde{Y}_2(s,\gamma^o)}{\bar{Y}_2(s,\gamma^o)}d\Lambda_{02}(s)\Big].$$

We deduce the asymptotic distribution of $\tilde{S}_n^*(t,\gamma^o)$ by considering the asymptotic distribution of each term in (4.19). For this end, consider the following theorem.

Theorem 4.5.2. *Given Assumptions 1-2 and under the null hypothesis,*

(1) *$\tilde{S}_n(t,\gamma^o)$ converges to a zero-mean normally distributed random variable with finite variance $\sigma^2_{\tilde{S}}(t)$, as n diverges to infinity, where*

$$\sigma^2_{\tilde{S}}(t) = \int_0^t W^2(s)\frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}_1(s)+\bar{y}_2(s)}d\Lambda_0(s). \tag{4.20}$$

(2) *$S_n^*(t,\gamma^o)$ converges to a zero-mean random variable with finite variance $\sigma^2_{S^*}(t)$ as n diverges to infinity.*

(3) *The two random variables $\tilde{S}_n(t,\gamma^o)$ and $S_n^*(t,\gamma^o)$ are uncorrelated.*

The proof of Theorem 4.5.2 is presented in Section 4.7.2. Summarizing the results of Conclusion 4.5.1 and Theorem 4.5.2, one can say that under the null hypothesis, our test statistic $S_n(t,\hat{\gamma})$ is asymptotically zero-mean normally distributed random variable, and its asymptotic variance can be approximated by $Var\{\tilde{S}_n(t,\gamma^o)\} + Var\{S_n^*(t,\gamma^o)\}$. Thus, based on direct calculations of the variances, as presented in Section 4.7.3, we present the following variance estimator of $S_n(t,\hat{\gamma})$

$$\begin{aligned}
\hat{\sigma}_I^2(t) &= \int_0^t \hat{W}_n^2(s) \sum_{j=1}^{2} \left\{ D_j^n(s,\hat{\gamma}) \right\}^2 \frac{d\bar{N}_j(s)}{\bar{Y}_j(s,\hat{\gamma})} \sum_{i=1}^{n} e^{\hat{\beta}^T Z_{ij}} Y_{ij}(s) \hat{E}(w_i) \\
&+ \int_0^t \int_0^t D_1^n(s,\hat{\gamma}) D_1^n(u,\hat{\gamma}) \sum_{i=1}^{n} Y_{i1}(s \vee u) e^{2\hat{\beta}^T Z_{i1}} \widehat{Var}(w_i \mid \mathcal{F}_{s\vee u-}) d\hat{\Lambda}_{01}(s) d\hat{\Lambda}_{01}(u) \\
&+ \int_0^t \int_0^t D_2^n(s,\hat{\gamma}) D_2^n(u,\hat{\gamma}) \sum_{i=1}^{n} Y_{i2}(s \vee u) e^{2\hat{\beta}^T Z_{i2}} \widehat{Var}(w_i \mid \mathcal{F}_{s\vee u-}) d\hat{\Lambda}_{02}(s) d\hat{\Lambda}_{02}(u) \\
&- 2 \int_0^t \int_0^t D_1^n(s,\hat{\gamma}) D_2^n(u,\hat{\gamma}) \sum_{i=1}^{n} Y_{i1}(s) Y_{i2}(u) e^{\hat{\beta}^T (Z_{i1}+Z_{i2})} \widehat{Var}(w_i \mid \mathcal{F}_{s\vee u-}) d\hat{\Lambda}_{01}(s) d\hat{\Lambda}_{02}(u).
\end{aligned} \tag{4.21}$$

For $\hat{E}(w_i)$ and $\widehat{Var}(w_i \mid \mathcal{F}_t)$ one can use $\hat{\gamma}$. Also, it should be noted that often $E(w_i)$ is set to be 1 for the model (4.1) to be identifiable. In these cases $\hat{E}(w_i) = 1$ $i = 1, ..., n$. However, as we show by extensive simulation study (Section 4.8), $Var\{S_n^*(t,\gamma^o)\}$ is of a negligible contribution to the total variance (less than 10%). Hence, we recommend to estimate the variance of the test statistic $S_n(t,\hat{\gamma})$ by the estimator of $Var\{\tilde{S}_n(t,\gamma^o)\}$. Specifically,

$$\hat{\sigma}_{II}^2(t) = \int_0^t \hat{W}_n^2(s) \sum_{j=1}^{2} \left\{ D_j^n(s,\hat{\gamma}) \right\}^2 d\hat{\Lambda}_{0j}(s) \sum_{i=1}^{n} e^{\hat{\beta}^T Z_{ij}} Y_{ij}(s) \hat{E}(w_i). \tag{4.22}$$

In conclusion, our proposed test statistic is defined by $S_n(t,\hat{\gamma})/\hat{\sigma}_n(t)$ and the rejection region corresponding to the null hypothesis (4.13) should be defined by the standard normal distribution.

4.5.3 Test for Equality of m Hazard Functions

Now we extend the test proposed in the previous section to test the null hypothesis (4.9) with $m > 2$ baseline hazard functions. Namely, we compare each of the m estimators of the cumulative baseline hazard functions $\left\{ \hat{\Lambda}_{0j} \right\}_{j=1}^{m}$ with an

estimator of the common cumulative baseline hazard function constructed under the null hypothesis. Let $\hat{\Lambda}_0$ be the estimated cumulative baseline hazard function under the null hypothesis (see [37] for details) in which the jump size of $\hat{\Lambda}_0$ at time s is defined by

$$\Delta\hat{\Lambda}_0(s) = \frac{\sum_{j=1}^{m} d\bar{N}_j(s)}{\bar{Y}_\cdot(s,\hat{\gamma})} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} dN_{ij}(s)}{\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{\psi}_i(s)Y_{ij}(s)e^{\hat{\beta}^T Z_{ij}}},$$

where $\bar{Y}_\cdot(s,\hat{\gamma}) = \sum_{j=1}^{m} \bar{Y}_j(s,\hat{\gamma})$.

We define $\mathbf{S_n}(\mathbf{t},\hat{\gamma}) = (S_{n1}(t,\hat{\gamma}),...,S_{nm}(t,\hat{\gamma}))^T$ to be the m-sample statistic. In the spirit of (4.14), we define

$$S_{nj}(t,\hat{\gamma}) = \frac{1}{\sqrt{n}}\int_0^t \hat{W}_{nj}(s)\frac{\bar{Y}_j(s,\hat{\gamma})\bar{Y}_\cdot(s,\hat{\gamma})}{\bar{Y}_j(s,\hat{\gamma})+\bar{Y}_\cdot(s,\hat{\gamma})}\Big\{d\hat{\Lambda}_{0j}(s) - d\hat{\Lambda}_0(s)\Big\} \quad j=1,...,m, \tag{4.23}$$

where $\hat{W}_{nj}(s)$ are nonnegative cadlag or caglad with total bounded variation. However, the special choice of weight processes such as

$$\hat{W}_{nj}(s) = \hat{W}_n(s)\frac{\bar{Y}_j(s,\hat{\gamma})+\bar{Y}_\cdot(s,\hat{\gamma})}{\bar{Y}_\cdot(s,\hat{\gamma})} \quad j=1,...,m,$$

where $\hat{W}_n(s)$ is nonnegative cadlag or caglad with total bounded variation, covers a wide variety of interesting cases (Andersen et al. [8], Section V.2). Hence, the above choice of weight process will be considered here. Then,

$$S_{nj}(t,\hat{\gamma}) = \frac{1}{\sqrt{n}}\int_0^t \hat{W}_n(s)\bar{Y}_j(s,\hat{\gamma})\Big\{d\hat{\Lambda}_{0j}(s) - d\hat{\Lambda}_0(s)\Big\} \qquad j=1,...,m, \tag{4.24}$$

and $\sum_{j=1}^{m} S_{nj}(t,\hat{\gamma}) = 0$. It is easy to verify that for $m=2$, $S_{n1}(t,\hat{\gamma})$ equals (4.14). Similar arguments used in the case of comparing two baseline hazard functions (Section 4.5.2) can be used here, such that we arrive to the following conclusion.

Conclusion 4.5.2. *An approximation of the asymptotic distribution of* $\mathbf{S_n}(\mathbf{t},\hat{\gamma})$ *is the asymptotic distribution of* $\tilde{\mathbf{S}}^*_{\mathbf{n}}(\mathbf{t},\gamma^{\mathbf{o}}) = \tilde{\mathbf{S}}_{\mathbf{n}}(\mathbf{t},\gamma^{\mathbf{o}}) + \mathbf{S}^*_{\mathbf{n}}(\mathbf{t},\gamma^{\mathbf{o}})$*, where the respective* j*-th components of* $\tilde{\mathbf{S}}_{\mathbf{n}}(\mathbf{t},\gamma^{\mathbf{o}})$ *and* $\mathbf{S}^*_{\mathbf{n}}(\mathbf{t},\gamma^{\mathbf{o}})$ *are*

$$\tilde{S}_{nj}(t,\gamma^o) = \int_0^t \frac{\hat{W}_n(s)}{\sqrt{n}}\bar{Y}_j(s,\gamma^o)\Big\{\frac{d\bar{M}_j(s)}{\bar{Y}_j(s,\gamma^o)} - \frac{d\bar{M}_\cdot(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}, \tag{4.25}$$

$$S^*_{nj}(t,\gamma^o) = \int_0^t \frac{\hat{W}_n(s)}{\sqrt{n}}\bar{Y}_j(s,\gamma^o)\Big\{\frac{\tilde{Y}_j(s,\gamma^o)d\Lambda_{0j}(s)}{\bar{Y}_j(s,\gamma^o)} - \frac{\tilde{Y}_\cdot(s,\gamma^o)d\Lambda_0(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}, \tag{4.26}$$

where $\bar{M}_\cdot(s) = \sum_{i=1}^{n} \bar{M}_j(s)$.

Note that given $w.$, $\bar{M}.(s)$ is a mean zero martingale with respect to $\mathcal{F}_{s-}$.

For the asymptotic distribution of $\tilde{\mathbf{S}}^*_{\mathbf{n}}(\mathbf{t}, \gamma^{\mathbf{o}})$ we present the following theorem and its proof can be found in Section 4.7.4.

Theorem 4.5.3. *Given Assumptions 1-2 and under the null hypothesis,*

(1) $\tilde{\mathbf{S}}_{\mathbf{n}}(\mathbf{t}, \gamma^{\mathbf{o}})$ *converges to a zero-mean multivariate normally distributed random variable with variance matrix* $\mathbf{V}(t)$ *and its* jk*-th component is defined by*

$$V_{jk}(t) = \begin{cases} \int_0^t W^2(s) \dfrac{\bar{y}_j(s) \sum_{r \neq j, r=1}^m \bar{y}_r(s)}{\bar{y}.(s)} \lambda_0(s) ds & k = j \\ -\int_0^t W^2(s) \dfrac{\bar{y}_j(s) \bar{y}_k(s)}{\bar{y}.(s)} \lambda_0(s) ds & k \neq j. \end{cases}$$

(2) $\mathbf{S}^*_{\mathbf{n}}(\mathbf{t}, \gamma^{\mathbf{o}})$ *converges to a zero-mean multivariate normal random variable with covariance matrix having finite diagonal entries and zero valued non-diagonal entries.*

(3) *The two random variables* $\tilde{\mathbf{S}}_{\mathbf{n}}(\mathbf{t}, \gamma^{\mathbf{o}})$ *and* $\mathbf{S}^*_{\mathbf{n}}(\mathbf{t}, \gamma^{\mathbf{o}})$ *are uncorrelated.*

Summarizing our results so far, we conclude that $S_n(t, \hat{\gamma})$ is asymptotically normal. Using similar arguments as for the case of testing equality of two hazard functions, motivates us to estimate the variance of $S_n(t, \hat{\gamma})$ based on the variance estimator of $\tilde{S}_n(t, \gamma^o)$. Hence our proposed estimator, denoted by $\hat{\mathbf{V}}(\mathbf{t})$ is given by

$$\begin{aligned} \hat{V}_{jj}(t) = \frac{1}{n} \int_0^t \hat{W}_n^2(s) \sum_{i=1}^n \Big[\Big\{ 1 - \frac{\bar{Y}_j(s, \hat{\gamma})}{\bar{Y}.(s, \hat{\gamma})} \Big\}^2 \hat{E}(w_i) Y_{ij}(s) e^{\hat{\beta}^T Z_{ij}} d\hat{\Lambda}_{0j}(s) \\ + \Big\{ \frac{\bar{Y}_j(s, \hat{\gamma})}{\bar{Y}.(s, \hat{\gamma})} \Big\}^2 \sum_{l \neq j}^m \hat{E}(w_i) Y_{il}(s) e^{\hat{\beta}^T Z_{il}} d\hat{\Lambda}_{0l}(s) \Big] \quad j = 1, ..., m. \end{aligned} \tag{4.27}$$

and for $k \neq j$

$$\begin{aligned} \hat{V}_{kj}(t) = \frac{1}{n} \int_0^t \hat{W}_n^2(s) \Big[\sum_{l \neq j,k} \frac{\bar{Y}_j(s, \hat{\gamma}) \bar{Y}_k(s, \hat{\gamma})}{\bar{Y}.^2(s, \hat{\gamma})} \hat{E}(w_i) Y_{il}(s) e^{\hat{\beta}^T Z_{il}} d\hat{\Lambda}_{0l}(s) \\ - \frac{\bar{Y}_j(s, \hat{\gamma})}{\bar{Y}.(s, \hat{\gamma})} \sum_{i=1}^n \hat{E}(w_i) Y_{ik}(s) e^{\hat{\beta}^T Z_{ik}} d\hat{\Lambda}_{0k}(s) \\ - \frac{\bar{Y}_k(s, \hat{\gamma})}{\bar{Y}.(s, \hat{\gamma})} \sum_{i=1}^n \hat{E}(w_i) Y_{ij}(s) e^{\hat{\beta}^T Z_{ij}} d\hat{\Lambda}_{0j}(s) \Big]. \end{aligned} \tag{4.28}$$

The details for the derivation of $\hat{\mathbf{V}}(\mathbf{t})$ are presented in Section 4.7.5. It is clear that $\mathbf{V}(\mathbf{t})$ has rank of $(m-1)$. Hence, we define $\hat{\mathbf{V}}^{\mathbf{o}}(\mathbf{t})$ as a $(m-1)\times(m-1)$ matrix obtained by deleting the last row and column of $\hat{\mathbf{V}}(\mathbf{t})$. Also, let $\mathbf{S}^{\mathbf{o}}_{\mathbf{n}}(\mathbf{t}) = \big(S_{n1}(t,\hat{\gamma}),...,S_{n(m-1)}(t,\hat{\gamma})\big)^T$. Then, our proposed test statistic is defined by $\mathbf{S}^{\mathbf{o}}_{\mathbf{n}}(\mathbf{t},\hat{\gamma})^T\Big[\hat{\mathbf{V}}^o(t)\Big]^{-1}\mathbf{S}^{\mathbf{o}}_{\mathbf{n}}(\mathbf{t},\hat{\gamma})$ and the rejection region should be defined by the $\chi^2(m-1)$ distribution.

It is clear that the above theory can be used directly for testing contrasts on the baseline hazard functions.

4.6 Sample Size Formula for Equality of Two Hazard Functions

In this section, we present a sample size formula under proportional means local alternative and certain simplifying assumptions for testing the equality of two baseline hazard functions. Specifically, let

$$H_1 : \quad \Lambda^n_{0j}(s) = \int_0^s \exp\{(-1)^{j-1}\varphi(u)/(2\sqrt{n})\}d\Lambda_0(u) \quad j=1,2 \quad \text{for all } s\in[0,\tau], \tag{4.29}$$

where Λ_0 is some unspecified cumulative hazard function with $\Lambda_0(s)<\infty$ and $\varphi(s)\neq 0$ for all $s\in[0,\tau]$. The above local alternative formulation was originally proposed by Kosorok and Lin [54] and also these alternatives can be found in the work of Gangnon and Kosorok [29].

It is easy to verify that the above $\Lambda^n_{0j}(s)$ $j=1,2$ satisfies the following assumptions:

5. For $j=1,2$

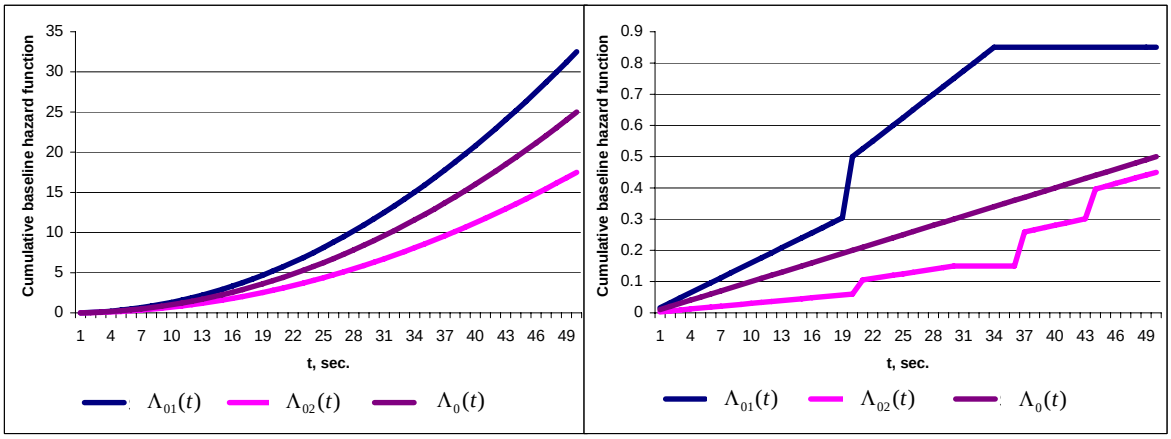
$$\sup_{s\in[0,\tau]} \Big| \, d\Lambda^n_{0j}(s)/d\Lambda_0(s) - 1 \, \Big| \to 0, \quad \text{as} \quad n\to\infty.$$

6. As $n\to\infty$,

$$\sup_{s\in[0,\tau]} \Big| \, \sqrt{n}\Big\{\frac{d\Lambda^n_{01}(s)}{d\Lambda^n_{02}(s)} - 1\Big\} - \varphi(s) \, \Big| \to 0$$

where φ is either cadlag or caglad with bounded total variation.

Figure 4.1 presents two examples of the cumulative baseline hazard functions under the above local alternatives defined by (4.29).



(a) (b)

Figure 4.1: An illustration of two possible alternatives satisfying definition (4.29).

Obviously, the family of local alternatives defined by (4.29) is much wider than the above two examples, however, the common structure is such that $\Lambda_{01}(t) < \Lambda_0(t) < \Lambda_{02}(t)$ (or vise versa) for all $t \in [0, \tau]$.

We start with the asymptotic distribution of $S_n(t, \gamma^o)$, under the local alternatives.

Theorem 4.6.1. *Given Assumptions 1 - 6, $S_n(t, \gamma^o)$ converges in distribution to a normal random variable with mean $\mu_1(t)$ and variance $\sigma^2(t)$, where*

$$\mu_1(t) = \int_0^t W(s)\varphi(s)\frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}_1(s) + \bar{y}_2(s)} d\Lambda_0(s) \tag{4.30}$$

and $\sigma^2(t) = \sigma^2_{\tilde{S}}(t)$ as defined in (4.20).

The proof of Theorem 4.6.1 is presented in Section 4.7.6.

Under the assumed contiguous alternative, we can approximate the power calculation as follows. For a fixed alternative set $\varphi(t) = \sqrt{n}\varphi^*(t)$. Then by (4.30) and the first order Taylor expansion we get $E_{H_1}\left\{S_n(t, \hat{\gamma})\right\} = \sqrt{n}\mu_1^*(t) + o(\sqrt{n})$, where

$$\mu_1^*(t) = \int_0^t W(s)\varphi^*(s)\frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}_1(s) + \bar{y}_2(s)} d\Lambda_0(s)\,. \tag{4.31}$$

Now, based on the limiting distribution of $S_n(t, \hat{\gamma})$, and under given significance level α and power π, we get

$$\begin{aligned}\pi &= P_{H_1}\left(\left|\frac{S_n(t, \hat{\gamma})}{\sigma(t)}\right| \geq Z_{1-\alpha/2}\right) \\ &= P_{H_1}\left(\left|\frac{S_n(t, \hat{\gamma}) - \sqrt{n}\mu_1^*(t)}{\sigma(t)} + \frac{\sqrt{n}\mu_1^*(t)}{\sigma(t)}\right| \geq Z_{1-\alpha/2}\right),\end{aligned}$$

where Z_p is the p-th quantile of the standard normal distribution. Then,

$$Z_{1-\alpha/2} - \frac{\sqrt{n}\mu_1^*(t)}{\sigma(t)} \approx -Z_\pi.$$

This gives us the following sample size formula

$$n = \frac{\left(Z_{1-\alpha/2} + Z_\pi\right)^2 \sigma^2(t)}{\{\mu_1^*(t)\}^2}. \tag{4.32}$$

However, in order to calculate the required sample size based on (4.32) one should estimate $\sigma^2(t)$ and $\mu_1^*(t)$ based on a pilot study or existing relevant datasets. In what follows, we propose simple estimators under simplifying assumptions, similar to those of [79]. These simple estimators provide a practical sample size formula.

Assume that the baseline hazard functions are continuous and the local alternatives satisfy $\varphi^*(s) = \varepsilon$ for all $s \in [0, \tau]$, when $\varepsilon \in \mathbb{R}$ and the weight function is constant $\hat{W}_n(s) \equiv 1$. We also assume that the limiting values of $\bar{Y}_j(s,\gamma)/n_j$ are $\pi_j(s)$, $j = 1, 2$ and the proportion of customers making the j-th call, n_j/n, converges to $p_j \in (0, 1]$, $j = 1, 2$. Then, based on Assumption 2, we replace $\bar{y}_j(s)$ by $p_j\pi_j(s)$. In addition, we assume that $\pi_1(s) = \pi_2(s) = \pi(s)$. Hence, (4.31) becomes

$$\begin{aligned} \mu_1^*(t) &= \int_0^t \varepsilon \frac{p_1 p_2 \pi^2(s)}{p_1\pi(s) + p_2\pi(s)} d\Lambda_0(s) \\ &= \varepsilon \frac{p_1 p_2}{p_1 + p_2} R(t), \end{aligned} \tag{4.33}$$

where $R(t) = \int_0^t \pi(s) d\Lambda_0(s)$. A simple estimator of $R(t)$ can be obtained as follows

$$\begin{aligned} \hat{R}(t) &= \int_0^t \{\hat{p}_1\hat{\pi}_1(s) d\hat{\Lambda}_{01}(s) + \hat{p}_2(s)\hat{\pi}_2 d\hat{\Lambda}_{02}(s)\} \\ &= \int_0^t \left\{ \frac{n_1}{n} \frac{\bar{Y}_1(s,\hat{\gamma})}{n_1} \frac{d\bar{N}_1(s)}{\bar{Y}_1(s,\hat{\gamma})} + \frac{n_2}{n} \frac{\bar{Y}_2(s,\hat{\gamma})}{n_2} \frac{d\bar{N}_2(s)}{\bar{Y}_2(s,\hat{\gamma})} \right\} \\ &= \frac{1}{n} \sum_{j=1}^{2} \bar{N}_j(t). \end{aligned} \tag{4.34}$$

Thus, a simplified sample size formula is given by

$$n = \frac{\left(Z_{1-\alpha/2} + Z_\pi\right)^2 \hat{\sigma}_{II}^2(t)}{\{\varepsilon \hat{p}_1 \hat{p}_2 \hat{R}(t)/(\hat{p}_1 + \hat{p}_2)\}^2}, \tag{4.35}$$

where $\hat{\sigma}^2_{II}(t)$ is given by (4.22).

Remark I. The widely used sample size formula proposed by Schoenfeld [76] is for the case of independent samples. By simulation study we show (Section 4.8) that Schoenfeld's formula underestimates the required sample size under dependent samples.

Remark II. Based on (4.20) and (4.30), and by the usual Cauchy-Schwartz argument, it can be shown that under the local alternative (4.29) the optimal weight function equals $W(s) = \varphi(s)$ for all $s \in [0, \tau]$.

4.7 Proofs

4.7.1 Proof of Theorem 4.5.1

Let

$$A_n(t, \hat{\gamma}) = \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s) \left\{ \frac{\bar{Y}_2(s, \hat{\gamma})}{\bar{Y}_\cdot(s, \hat{\gamma})} dM_1(s) - \frac{\bar{Y}_1(s, \hat{\gamma})}{\bar{Y}_\cdot(s, \hat{\gamma})} dM_2(s) \right\}$$

and write $S_n(t, \hat{\gamma}) = A_n(t, \hat{\gamma}) + S^{**}_n(t)$. The first order Taylor expansion of $A_n(t, \hat{\gamma})$ about γ^o gives

$$A_n(t, \hat{\gamma}) \approx \tilde{S}_n(t, \gamma^o) + \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s) \left\{ \mathbf{Q}_2^T(s, \gamma^o) dM_1(s) - \mathbf{Q}_1^T(s, \gamma^o) dM_2(s) \right\} (\hat{\gamma} - \gamma^o). \tag{4.36}$$

Since $M_j(t)/\sqrt{n}$ converges in distribution as $n \to \infty$ (it is asymptotically normal given $w_\cdot$) and using Assumptions 1 and 3-4 stating the existence of deterministic functions $W(s)$ and $g_{lj}(s)$, $l = 1, ..., p+1$ $j = 1, ..., m$, such that

$$\sup_{s \in [0,\tau]} \mid Q_{lj}(s, \gamma^o) - g_{lj}(s) \mid \to 0 \quad \sup_{s \in [0,\tau]} \left\{ \hat{W}_n(s) - W(s) \right\} \longrightarrow 0,$$

we get that the conditional distribution of

$$B_n(t, \gamma^o) = \frac{1}{\sqrt{n}} \int_0^t \hat{W}_n(s) \left\{ \mathbf{Q}_2^T(s, \gamma^o) dM_1(s) - \mathbf{Q}_1^T(s, \gamma^o) dM_2(s) \right\},$$

conditioning on $w_\cdot$, convergence to zero-mean multivariate normally distributed random variable with finite entries of the covariance matrix that are free of the frailties. Hence, this is also the unconditional asymptotic distribution of $B_n(t, \gamma^o)$. Then, given Claim 3.3, the second term of (4.36) goes to zero as $n \to \infty$, by Slutsky's theorem.

4.7.2 Proof of Theorem 4.5.2

We present the proofs of each theorem's statements in the sequence.

Proof of statement (1). Given $w.$, $\tilde{S}_n(t,\gamma^o)$ is a mean-zero martingale. Hence, to show that given $w.$ it converges to a normally distributed random variable, one needs to show that the conditions of the martingale central limit theorem (see [1], Section 2.3.3, for details) hold. Namely,

(i) $\sum_{j=1}^{2}\left\{D_j^n(s,\gamma^o)\right\}^2\lambda_j^n(s,\gamma^o) \longrightarrow_p v(s) \qquad$ for all $\quad s\in[0,\tau], \quad$ as $\quad n\to\infty,$

where $\lambda_j^n(s,\gamma^o)=\sum_{i=1}^{n} Y_{ij}(s)\tilde{h}_{ij}(s,\beta^o)\lambda_{0j}(s)$ is a sum of intensity processes of n independent customers, $\tilde{h}_{ij}(s,\beta^o)=w_i e^{\beta^{oT}Z_{ij}}$ and $V(t)=\int_0^t v(s)ds$ is the variance of the limiting process.

(ii) $D_j^n(s,\gamma^o)\longrightarrow_p 0 \qquad$ for all $\ j=1,2\ $ and $\ s\in[0,\tau],\ $ as $n\to\infty.$

In our case, under the null hypothesis $\lambda_{0j}(s)=\lambda_0(s)$, $j=1,2$ for all $s\in[0,\tau]$. Therefore, under the null hypothesis and Assumptions 1 - 2, we obtain

$$\begin{aligned}
&\sum_{j=1}^{2}\left\{D_j^n(s,\gamma^o)\right\}^2\lambda_j^n(s,\gamma^o)= \\
&=\sum_{j=1}^{2}\left\{\frac{\hat{W}_n(s)}{\sqrt{n}}\frac{\bar{Y}_{3-j}(s,\gamma^o)}{\bar{Y}_1(s,\gamma^o)+\bar{Y}_2(s,\gamma^o)}\right\}^2\sum_{i=1}^{n}Y_{ij}(s)\tilde{h}_{ij}(s,\gamma^o)\lambda_0(s) \\
&=\frac{\hat{W}_n^2(s)}{n}\frac{\bar{Y}_2^2(s,\gamma^o)\tilde{Y}_1(s,\gamma^o)+\bar{Y}_1^2(s,\gamma^o)\tilde{Y}_2(s,\gamma^o)}{\{\bar{Y}_1(s,\gamma^o)+\bar{Y}_2(s,\gamma^o)\}^2}\lambda_0(s) \\
&\longrightarrow_p W^2(s)\frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}_1(s)+\bar{y}_2(s)}\lambda_0(s), \qquad as \quad n\to\infty,
\end{aligned} \tag{4.37}$$

and

$$D_j^n(s,\gamma^o)=\frac{1}{\sqrt{n}}\hat{W}_n(s)\frac{\bar{Y}_{3-j}(s,\gamma^o)/n}{\{\bar{Y}_1(s,\gamma^o)+\bar{Y}_2(s,\gamma^o)\}/n}\longrightarrow_p 0, \qquad as \quad n\to\infty.$$

Hence, we conclude that $\tilde{S}_n(t,\gamma^o)$, given $w.$, converges to a normally distributed random variable with moments that are free of the frailties $w.$. Therefore, $\tilde{S}_n(t,\gamma^o)$ also converges to a normally distributed random variable with the same parameters.

Proof of statement (2). Note that $S_n^*(t,\gamma^o)$ can be rewritten in the following form

$$\begin{aligned} S_n^*(t,\gamma^o) &= \frac{1}{\sqrt{n}}\int_0^t \hat{W}_n(s)\frac{\bar{Y}_1(s,\gamma^o)\bar{Y}_2(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\{\frac{\tilde{Y}_1(s,\gamma^o)}{\bar{Y}_1(s,\gamma^o)} - \frac{\tilde{Y}_2(s,\gamma^o)}{\bar{Y}_2(s,\gamma^o)}\Big\}d\Lambda_0(s) \\ &= \int_0^t \hat{W}_n(s)\Big(\frac{\bar{Y}_2(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\{\bar{M}_1(s) - \bar{M}_1^*(s)\Big\} \\ &- \frac{\bar{Y}_1(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\{\bar{M}_1(s) - \bar{M}_1^*(s)\Big\}\Big)d\Lambda_0(s), \end{aligned}$$

where $\bar{M}_j^* = \sum_{i=1}^n M_{ij}^*(s)$ is a mean-zero martingale of the process $\bar{N}_j(s)$. Then, applying the martingale central limit theorem in a way analogous to the proof of statement (1), we obtain that $S_n^*(t,\gamma^o)$ is asymptotically normally distributed.

Obviously, that $S_n^*(t,\gamma^o)$ has mean zero and for the simplicity of its variance calculation we let

$$g_j(s) = \hat{W}_n(s)\lambda_0(s)\frac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)} \quad \text{and} \quad X_j^*(s) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_{ij}(s)e^{\beta^T Z_{ij}}\Big\{w_i - E(w_i \mid \mathcal{F}_{s-})\Big\}$$

for $j = 1, 2$. Then,

$$\begin{aligned} Var\{S_n^*(t,\gamma^o)\} &= Var\Big\{\int_0^t g_1(s)X_1^*(s)ds\Big\} + Var\Big\{\int_0^t g_2(s)X_2^*(s)ds\Big\} \\ &- 2Cov\Big\{\int_0^t g_1(s)X_1^*(s)ds, \int_0^t g_2(s)X_2^*(s)ds\Big\}. \end{aligned} \tag{4.38}$$

Since $X_j^*(s)$ $j = 1, 2$ have mean zero, by using the law of total expectation we get

$$\begin{aligned} &Var\Big\{\int_0^t g_j(s)X_j^*(s)ds\Big\} = \\ &= E\left(E\Big\{\int_0^t g_j(s)X_j^*(s)ds\int_0^t g_j(u)X_j^*(u)du \mid \mathcal{F}_{s\vee u-}\Big\}\right) \\ &= \frac{1}{n}E\Big\{\int_0^t\int_0^t g_j(s)g_j(u)\sum_{i=1}^n Y_{ij}(s\vee u)e^{2\beta^T Z_{ij}}Var(w_i \mid \mathcal{F}_{s\vee u-})\Big\}dsdu, \end{aligned} \tag{4.39}$$

and

$$\begin{aligned} &Cov\Big\{\int_0^t g_1(s)X_1^*(s)ds, \int_0^t g_2(s)X_2^*(s)ds\Big\} = \\ &= \frac{1}{n}E\Big\{\int_0^t\int_0^t g_1(s)g_2(u)\sum_{i=1}^n Y_{i1}(s)Y_{i2}(u)e^{\beta^T(Z_{i1}+Z_{i2})}Var(w_i \mid \mathcal{F}_{s\vee u-})\Big\}dsdu. \end{aligned} \tag{4.40}$$

Combining (4.38)-(4.40) we get

$$
\begin{aligned}
&Var\{S_n^*(t,\gamma^o)\} = \\
&= \frac{1}{n}\Big(\sum_{j=1}^{m} E\Big\{\int_0^t\int_0^t g_j(s)g_j(u)\sum_{i=1}^{n} Y_{ij}(s\vee u)e^{2\beta^T Z_{ij}} Var(w_i \mid \mathcal{F}_{s\vee u-})\Big\}dsdu \\
&- 2E\Big\{\int_0^t\int_0^t g_1(s)g_2(u)\sum_{i=1}^{n} Y_{i1}(s)Y_{i2}(u)e^{\beta^T(Z_{i1}+Z_{i2})} Var(w_i \mid \mathcal{F}_{s\vee u-})\Big\}dsdu\Big).
\end{aligned}
\tag{4.41}
$$

Hence, it is easy to see that $Var\Big\{S_n^*(t,\gamma^o)\Big\} < \infty$.

Proof of statement (3). Note that under the null hypothesis, the covariance between $\tilde{S}_n(t,\gamma^o)$ and $S_n^*(t,\gamma^o)$ can be written as follows

$$
\begin{aligned}
&Cov\Big(\tilde{S}_n(t,\gamma^o), S_n^*(t,\gamma^o)\Big) \\
&= E\Big[\int_0^t D_n(s,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\} \\
&\int_0^t D_n(s,\gamma^o)\Big\{\frac{\tilde{Y}_1(s,\gamma^o)d\Lambda_{01}(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{\tilde{Y}_2(s,\gamma^o)d\Lambda_{02}(s)}{\bar{Y}_2(s,\gamma^o)}\Big\}\Big] \\
&= \int_0^t\int_0^t E\Big[D_n(s,\gamma^o)D_n(u,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\}\Big\{\frac{\tilde{Y}_1(u,\gamma^o)}{\bar{Y}_1(u,\gamma^o)} - \frac{\tilde{Y}_2(u,\gamma^o)}{\bar{Y}_2(u,\gamma^o)}\Big\}d\Lambda_0(u)\Big].
\end{aligned}
\tag{4.42}
$$

Now we show that $Cov\{\tilde{S}_n(t,\gamma^o), S_n^*(t,\gamma^o)\} = 0$ by showing that given w.

$$
E\left[D_n(s,\gamma^o)D_n(u,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\}\Big\{\frac{\tilde{Y}_1(u,\gamma^o)}{\bar{Y}_1(u,\gamma^o)} - \frac{\tilde{Y}_2(u,\gamma^o)}{\bar{Y}_2(u,\gamma^o)}\Big\}d\Lambda_0(u)\right] = 0,
$$

for all s, $u \in [0,\tau]$. Indeed, for $s \geq u$ we get

$$
\begin{aligned}
&E\Big[E\Big(D_n(s,\gamma^o)D_n(u,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\} \\
&\qquad\qquad \Big\{\frac{\tilde{Y}_1(u,\gamma^o)}{\bar{Y}_1(u,\gamma^o)} - \frac{\tilde{Y}_2(u,\gamma^o)}{\bar{Y}_2(u,\gamma^o)}\Big\}d\Lambda_0(u) \mid \mathcal{F}_{s-}\Big)\Big] \\
&= E\Big[D_n(s,\gamma^o)D_n(u,\gamma^o)\Big\{\frac{\tilde{Y}_1(u,\gamma^o)}{\bar{Y}_1(u,\gamma^o)} - \frac{\tilde{Y}_2(u,\gamma^o)}{\bar{Y}_2(u,\gamma^o)}\Big\} \\
&\qquad\qquad E\Big(\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\} \mid \mathcal{F}_{s-}\Big)d\Lambda_0(u)\Big] = 0,
\end{aligned}
$$

because $E\left(\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)} \mid \mathcal{F}_{s-}\right) = 0$ given $w.$, as an expectation of zero mean martingales. For $u > s$, we get

$$\begin{aligned}
&E\Big[E\Big(D_n(s,\gamma^o)D_n(u,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\} \\
&\qquad\qquad \Big\{\frac{\tilde{Y}_1(u,\gamma^o)}{\bar{Y}_1(u,\gamma^o)} - \frac{\tilde{Y}_2(u,\gamma^o)}{\bar{Y}_2(u,\gamma^o)}\Big\}d\Lambda_0(u) \mid \mathcal{F}_{u-}\Big)\Big] \\
&= E\Big[D_n(s,\gamma^o)D_n(u,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\} \\
&\qquad\qquad E\Big(\Big\{\frac{\tilde{Y}_1(u,\gamma^o)}{\bar{Y}_1(u,\gamma^o)} - \frac{\tilde{Y}_2(u,\gamma^o)}{\bar{Y}_2(u,\gamma^o)}\Big\} \mid \mathcal{F}_{u-}\Big)d\Lambda_0(u)\Big] = 0.
\end{aligned}$$

4.7.3 An Estimator of the Variance of $S_n(t,\hat{\gamma})$

In the following, we generate our variance estimators of $Var\{\tilde{S}_n(t,\gamma^o)\}$ and $Var\{S_n^*(t,\gamma^o)\}$. Let us start with an estimator of $Var\{\tilde{S}_n(t,\gamma^o)\}$. By using the law of total variance we get

$$Var\{\tilde{S}_n(t,\gamma^o)\} = E\left[Var\{\tilde{S}_n(t,\gamma^o) \mid w.\}\right] + Var\left[E\{\tilde{S}_n(t,\gamma^o) \mid w.\}\right].$$

It is clear that $\tilde{S}_n(t,\gamma^o)$ given $w.$ is a mean-zero martingale under the null hypothesis, as a difference between two mean zero martingales. Hence,

$$Var\{\tilde{S}_n(t,\gamma^o)\} = E\left(Var\left[\int_0^t D_n(s,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\} \mid w.\right]\right).$$

Since calls of customer i are conditionally independent given w_i, the predictable variation process of $\tilde{S}_n(t,\gamma^o)$, given $w.$, is given by

$$\begin{aligned}
&< \tilde{S}_n \mid w. > (t,\gamma^o) = \int_0^t D_n^2(s,\gamma^o)Var\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} - \frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)} \mid w., \mathcal{F}_{s-}\Big\} \\
&= \int_0^t D_n^2(s,\gamma^o)\left(Var\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)} \mid w., \mathcal{F}_{s-}\Big\} + Var\Big\{\frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)} \mid w., \mathcal{F}_{s-}\Big\}\right).
\end{aligned}$$

Since $Var\big\{dM_{ij}(s) \mid w., \mathcal{F}_{s-}\big\} = Y_{ij}(s)e^{\beta^T Z_{ij}}w_i\lambda_{0j}(s)ds$ we get

$$< \tilde{S}_n \mid w. > (t,\gamma^o) = \sum_{i=1}^n w_i \int_0^t D_n^2(s,\gamma^o)\Big\{\frac{e^{\beta^T Z_{i1}}Y_{i1}(s)}{\bar{Y}_1^2(s,\gamma^o)}d\Lambda_{01}(s) + \frac{e^{\beta^T Z_{i2}}Y_{i2}(s)}{\bar{Y}_2^2(s,\gamma^o)}d\Lambda_{02}(s)\Big\}.$$

Then, the expectation with respect to the unknown frailties gives

$$\sum_{i=1}^n E(w_i)\int_0^t D_n^2(s,\gamma^o)\Big\{\frac{e^{\beta^T Z_{i1}}Y_{i1}(s)}{\bar{Y}_1^2(s,\gamma^o)}d\Lambda_{01}(s) + \frac{e^{\beta^T Z_{i2}}Y_{i2}(s)}{\bar{Y}_2^2(s,\gamma^o)}d\Lambda_{02}(s)\Big\}. \quad (4.43)$$

The variance of $Var\{S_n^*(t,\gamma^o)\}$ is presented in (4.41). Therefore, we replace all the unknown parameters in (4.43) and (4.41) by their estimates from Section 4.4 and get the estimators as presented in (4.21) and (4.22).

4.7.4 Proof of Theorem 4.5.3

Proof of statement (1). We start by rewriting each $\tilde{S}_{nj}(t,\gamma^o)$ as follows

$$\tilde{S}_{nj}(t,\gamma^o) = \int_0^t \frac{1}{\sqrt{n}}\hat{W}_n(s)\Big\{d\bar{M}_j(s) - \frac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}d\bar{M}_\cdot(s)\Big\}$$

and we show that given $w_\cdot$, the sequence $\tilde{\mathbf{S}}_\mathbf{n}(t,\gamma^o)$ converges to m-variate zero-mean Gaussian random variable. Let $\mathbf{M^{(n)}}(s) = (\bar{M}_1(s),...,\bar{M}_m(s))^T$. Then, based on the martingale central limit theorem, it is enough to show that the following conditions hold (see [1], Appendix B.3, for details)

$$(i)\quad \langle\int_0^t \mathbf{D^{(n)}}d\mathbf{M^{(n)}} \;\big|\; w_\cdot\rangle(s,\gamma^o) = \int_0^t \mathbf{D^{(n)}}(s,\gamma^o)Var\Big\{d\mathbf{M^{(n)}}(s) \;\big|\; \mathcal{F}_{s-}, w_\cdot\Big\}\mathbf{D^{(n)}}^T(s,\gamma^o)$$

such that

$$\mathbf{D^{(n)}}(s,\gamma^o)Var\Big\{d\mathbf{M^{(n)}}(s) \;\big|\; \mathcal{F}_{s-}, w_\cdot\Big\}\mathbf{D^{(n)}}^T(s,\gamma^o) \longrightarrow_p \mathbf{v}(s)$$

for all $s\in[0,t]$ as $n\to\infty$ and $\mathbf{V}(t) = \int_0^t \mathbf{v}(s)ds$.

Here $\mathbf{D^{(n)}}(s,\gamma^o)$ is a $m\times m$ matrix whose (k,j) entry equals

$$D_{jk}^{(n)}(s,\gamma^o) = \begin{cases} \dfrac{\hat{W}_n(s)}{\sqrt{n}}\Big\{1 - \dfrac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\} & k=j \\ -\dfrac{\hat{W}_n(s)}{\sqrt{n}}\dfrac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)} & k\neq j \end{cases}$$

(ii) $D_{jk}^{(n)}(s,\gamma^o) \longrightarrow_p 0 \quad j,k=1,...,m \;\; s\in[0,\tau]$, as $n\to\infty$.

Indeed, under Assumptions 1-2 and given the null hypothesis, $V_{jk}^{(n)}(s)$, the (j,k) component of the integrand of $\langle\int_0^t \mathbf{D^{(n)}}d\mathbf{M^{(n)}} \;\big|\; w_\cdot\rangle(s,\gamma^o)$ converges as follows

$$\begin{aligned} V_{jj}^{(n)}(t) &= \frac{1}{n}\int_0^t \hat{W}_n^2(s)\Big(\Big\{1 - \frac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}^2\tilde{Y}_j(s,\gamma^o)\lambda_{0j}(s) \\ &\quad + \frac{\bar{Y}_j^2(s,\gamma^o)}{\bar{Y}_\cdot^2(s,\gamma^o)}\sum_{r\neq j,r=1}^m \tilde{Y}_r(s,\gamma^o)\lambda_{0r}(s)\Big)ds \\ &\longrightarrow_p \int_0^t W^2(s)\frac{\bar{y}_j(s)\sum_{r\neq j,r=1}^m \bar{y}_r(s)}{\bar{y}_\cdot(s)}\lambda_0(s)ds \quad \text{as} \quad n\to\infty, \quad j=1,...,m \end{aligned}$$

and for $j \neq k$, $j,k = 1,...,m$

$$V_{jk}^{(n)}(t) = \frac{1}{n}\int_0^t \hat{W}_n^2(s)\Big\{\frac{\bar{Y}_j(s,\gamma^o)\bar{Y}_k(s,\gamma^o)}{\bar{Y}_\cdot^2(s,\gamma^o)}\sum_{r=1}^m \tilde{Y}_r(s,\gamma^o)\lambda_{0r}(s)$$
$$- \frac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\tilde{Y}_k(s,\gamma^o)\lambda_{0k}(s) - \frac{\bar{Y}_k(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\tilde{Y}_j(s,\gamma^o)\lambda_{0j}(s)\Big\}ds$$
$$\longrightarrow_p - \int_0^t W^2(s)\frac{\bar{y}_j(s)\bar{y}_k(s)}{\bar{y}_\cdot(s)}\lambda_0(s)ds \quad \text{as} \quad n\to\infty.$$

For Condition (ii), it is easy to see that under Assumptions 1-2, for $j = 1,...,m$, since

$$\frac{\bar{Y}_j(s,\gamma^0)}{\bar{Y}_\cdot(s,\gamma^0)} \longrightarrow \frac{\bar{y}_j(s,\gamma^0)}{\bar{y}_\cdot(s,\gamma^0)}, \qquad as \quad n\to\infty,$$

$$D_{jj}^{(n)}(s,\gamma^o) = \frac{\hat{W}_n(s)}{\sqrt{n}}\Big\{1 - \frac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\} \longrightarrow_p 0, \qquad as \quad n\to\infty$$

and

$$D_{jk}^{(n)}(s,\gamma^o) = \frac{\hat{W}_n(s)}{\sqrt{n}}\frac{\bar{Y}_j(s,\gamma^o)/n}{\bar{Y}_\cdot(s,\gamma^o)/n} \longrightarrow_p 0, \qquad as \quad n\to\infty.$$

As before, since the conditional asymptotic distribution of $\tilde{\mathbf{S}}_\mathbf{n}(\mathbf{t},\gamma^\mathbf{o})$ given the frailty variates is free of the frailties $w_\cdot$ we conclude that this is also the asymptotic distribution of $\tilde{\mathbf{S}}_\mathbf{n}(t,\gamma^o)$.

Proof of statement (2). Since for $j = 1,...,m$

$$E\Big\{\frac{\tilde{Y}_j(s,\gamma^o)}{\bar{Y}_j(s,\gamma^o)} - \frac{\tilde{Y}_\cdot(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)} \mid \mathcal{F}_{s-}\Big\} =$$
$$= \frac{\sum_{i=1}^n Y_{ij}(s)e^{\beta Z_{ij}}E\left[w_i \mid \mathcal{F}_{s-}\right]}{\bar{Y}_j(s,\gamma^o)} - \frac{\sum_{k=1}^m\sum_{i=1}^n Y_{ik}(s)e^{\beta Z_{ik}}E\left[w_i \mid \mathcal{F}_{s-}\right]}{\bar{Y}_2(s,\gamma^o)} \tag{4.44}$$
$$= \frac{\bar{Y}_j(s,\gamma^o)}{\bar{Y}_j(s,\gamma^o)} - \frac{\bar{Y}_\cdot(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)} = 0$$

it is easy to show that under the null hypothesis

$$E\Big[S_{nj}^*(t,\gamma^o)\Big] = E\Big[\int_0^t \frac{\hat{W}_n(s)}{\sqrt{n}}\bar{Y}_j(s,\gamma^o)\Big\{\frac{\tilde{Y}_j(s,\gamma^o)}{\bar{Y}_j(s,\gamma^o)} - \frac{\tilde{Y}_\cdot(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}d\Lambda_0(s)\Big] = 0, \tag{4.45}$$

for $j = 1,...,m$. Also, using again the law of total expectation by conditioning

on $\mathcal{F}_{u\vee s-}$ we obtain that under the null hypothesis

$$
\begin{aligned}
&Cov\Big[S^*_{nj}(t,\gamma^o), S^*_{nk}(t,\gamma^o)\Big] = \\
&= \int_0^\tau \int_0^t E\Big[\frac{\hat{W}_n(s)}{\sqrt{n}}\bar{Y}_j(s,\gamma^o)\frac{\hat{W}_n(u)}{\sqrt{n}}\bar{Y}_k(u,\gamma^o)\Big\{\frac{\tilde{Y}_j(s,\gamma^o)}{\bar{Y}_j(s,\gamma^o)} - \frac{\tilde{Y}_\cdot(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}d\Lambda_0(s) \\
&\Big\{\frac{\tilde{Y}_k(u,\gamma^o)}{\bar{Y}_k(u,\gamma^o)} - \frac{\tilde{Y}_\cdot(u,\gamma^o)}{\bar{Y}_\cdot(u,\gamma^o)}\Big\}d\Lambda_0(u)\Big] = 0.
\end{aligned}
$$

Using similar arguments as in the proof of Theorem 4.5.2, one can show that each of the j-th components of $\tilde{S}^*_{nj}(t,\gamma^o)$ is asymptotically normally distributed with a finite variance.

Proof of statement (3). Note that under the null hypothesis, the covariance between $\tilde{S}_{nj}(t,\gamma^o)$ and $S^*_{nk}(t,\gamma^o)$ for all $j,k = 1,...,m$ can be written as follows

$$
\begin{aligned}
&Cov\Big\{\tilde{S}_{nj}(t,\gamma^o), S^*_{nk}(t,\gamma^o)\Big\} \\
&= E\Big[\int_0^t \frac{\hat{W}_n(s)}{\sqrt{n}}\bar{Y}_j(s,\gamma^o)\Big\{\frac{d\bar{M}_j(s)}{\bar{Y}_j(s,\gamma^o)} - \frac{d\bar{M}_\cdot(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\} \\
&\int_0^t \frac{\hat{W}_n(s)}{\sqrt{n}}\bar{Y}_k(s,\gamma^o)\Big\{\frac{\tilde{Y}_k(s,\gamma^o)}{\bar{Y}_k(s,\gamma^o)} - \frac{\tilde{Y}_\cdot(s,\gamma^o)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}d\Lambda_0(s)\Big] \qquad (4.46)\\
&= \int_0^\tau \int_0^t E\Big[\frac{\hat{W}_n(s)}{\sqrt{n}}\bar{Y}_j(s,\gamma^o)\frac{\hat{W}_n(u)}{\sqrt{n}}\bar{Y}_k(u,\gamma^o)\Big\{\frac{d\bar{M}_j(s)}{\bar{Y}_j(s,\gamma^o)} - \frac{d\bar{M}_\cdot(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\} \\
&\Big\{\frac{\tilde{Y}_k(u,\gamma^o)}{\bar{Y}_k(u,\gamma^o)} - \frac{\tilde{Y}_\cdot(u,\gamma^o)}{\bar{Y}_\cdot(u,\gamma^o)}\Big\}\Big]d\Lambda_0(u).
\end{aligned}
$$

Now, let us show that $Cov\Big\{\tilde{S}_{nj}(t,\gamma^o), S^*_{nk}(t,\gamma^o)\Big\} = 0$ by showing that

$$
E\left[\frac{\hat{W}_n(s)\hat{W}_n(u)}{n}\bar{Y}_j(s,\gamma^o)\bar{Y}_k(u,\gamma^o)\Big\{\frac{d\bar{M}_j(s)}{\bar{Y}_j(s,\gamma^o)} - \frac{d\bar{M}_\cdot(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}\Big\{\frac{\tilde{Y}_k(u,\gamma^o)}{\bar{Y}_k(u,\gamma^o)} - \frac{\tilde{Y}_\cdot(u,\gamma^o)}{\bar{Y}_\cdot(u,\gamma^o)}\Big\}\right] = 0,
$$

for all u, $s \in [0,\tau]$. Indeed, for $s \geq u$ we have

$$
\begin{aligned}
E\Big[\frac{\hat{W}_n(s)\hat{W}_n(u)}{n}\bar{Y}_j(s,\gamma^o)\bar{Y}_k(u,\gamma^o)\Big\{\frac{\tilde{Y}_k(u,\gamma^o)}{\bar{Y}_k(u,\gamma^o)} - \frac{\tilde{Y}_\cdot(u,\gamma^o)}{\bar{Y}_\cdot(u,\gamma^o)}\Big\} \\
E\Big(\Big\{\frac{d\bar{M}_j(s)}{\bar{Y}_j(s,\gamma^o)} - \frac{d\bar{M}_\cdot(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\} \mid \mathcal{F}_{s-}\Big)\Big] = 0,
\end{aligned}
$$

because $E\left(\Big\{\frac{d\bar{M}_j(s)}{\bar{Y}_j(s,\gamma^o)} - \frac{d\bar{M}_\cdot(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\} \mid \mathcal{F}_{s-}\right) = 0$, as an expectation of zero mean

martingales given $w_\cdot$. For $u > s$ we obtain

$$E\Big[\frac{\hat{W}_n(s)\hat{W}_n(u)}{n}\bar{Y}_j(s,\gamma^o)\bar{Y}_k(u,\gamma^o)\Big\{\frac{d\bar{M}_j(s)}{\bar{Y}_j(s,\gamma^o)}-\frac{d\bar{M}_\cdot(s)}{\bar{Y}_\cdot(s,\gamma^o)}\Big\}$$
$$E\Big(\Big\{\frac{\tilde{Y}_k(u,\gamma^o)}{\bar{Y}_k(u,\gamma^o)}-\frac{\tilde{Y}_\cdot(u,\gamma^o)}{\bar{Y}_\cdot(u,\gamma^o)}(u)\Big\}\mid\mathcal{F}_{u-}\Big)\Big]=0.$$

4.7.5 The estimation of $\hat{\mathbf{V}}(t)$

Calls of customer i given $w_\cdot$ are independent. Therefore, the predictable variation process of $\tilde{\mathbf{S}}_\mathbf{n}(t,\gamma^o)$ given $w_\cdot$ is given by

$$<\tilde{S}_{nj}\mid w_\cdot>(t,\gamma^o)=V_{jj}^{(n)}(t)\quad\text{and}\quad<\tilde{S}_{nk},\tilde{S}_{nj}\mid w_\cdot>(t,\gamma^o)=V_{jk}^{(n)}(t)$$

for $j,k=1,...,m$ since $Var\{d\bar{M}_j(s)\mid w_\cdot,\mathcal{F}_{t-}\}=\tilde{Y}_j(s,\gamma)\lambda_{0j}(s)ds$. Then, by taking the expectation with respect to $w_\cdot$ we obtain

$$V_{jj}^{(n)}(t)=\sum_{i=1}^{n}\sum_{k=1}^{m}\int_0^t\Big\{D_{jk}^{(n)}(s,\gamma^o)\Big\}^2E(w_i)e^{\beta^TZ_{ik}}Y_{ik}(s)d\Lambda_{0k}(s)\quad j=1,...,m \tag{4.47}$$

and for $j\neq k$, $j,k=1,...,m$

$$\begin{aligned}V_{jk}^{(n)}=\sum_{i=1}^{n}\Big[&\sum_{l=1}^{m}\int_0^t D_{lj}^{(n)}(s,\gamma^o)D_{kl}^{(n)}(s,\gamma^o)Y_{il}(s)e^{\beta^TZ_{il}}d\Lambda_{0l}(s)\\&-\int_0^\tau D_{jk}^{(n)}(s,\gamma^o)E(w_i)e^{\beta^TZ_{ik}}Y_{ik}(s)d\Lambda_{0k}(s)\\&-\int_0^\tau D_{kj}^{(n)}(s,\gamma^o)E(w_i)e^{\beta^TZ_{ij}}Y_{ij}(s)d\Lambda_{0j}(s)\Big].\end{aligned}$$

Finally, by replacing all the unknown parameters by their estimates we obtain the estimators (4.27) and (4.28).

4.7.6 Proof of Theorem 4.6.1

Write

$$\begin{aligned}S_n(t,\gamma^o)=&\frac{1}{\sqrt{n}}\int_0^t\hat{W}_n(s)\mathcal{G}(s,\gamma^o)\Big\{\frac{d\bar{M}_1(s)}{\bar{Y}_1(s,\gamma^o)}-\frac{d\bar{M}_2(s)}{\bar{Y}_2(s,\gamma^o)}\Big\}\\&+\frac{1}{n}\int_0^t\hat{W}_n(s)\mathcal{G}(s,\gamma^o)\frac{\tilde{Y}_2(s,\gamma^o)}{\bar{Y}_2(s,\gamma^o)}d\Lambda_{02}^n(s)\sqrt{n}\Big\{\frac{\tilde{Y}_1(s,\gamma^o)\bar{Y}_2(s,\gamma^o)d\Lambda_{01}^n(s)}{\tilde{Y}_2(s,\gamma^o)\bar{Y}_1(s,\gamma^o)d\Lambda_{02}^n(s)}-1\Big\}.\end{aligned} \tag{4.48}$$

First, we show that the first term of the right-hand side of (4.48) converges to a normal random variable with mean zero. The proof is similar to that of Theorem 4.5.2. Hence, we need to show that Conditions (i) and (ii) of Section 4.7.2 hold. Proof of Condition (ii) is exactly the same as in the proof of Theorem 4.5.2 and the proof of Condition (i) is slightly different, as follows. Using Assumptions 1, 2 and 5, we have

$$\sum_{j=1}^{2}\left\{D_j^n(s,\gamma^o)\right\}^2\lambda_j^n(s)=\sum_{j=1}^{2}\sum_{i=1}^{n}\left\{\frac{\hat{W}_n(s)}{\sqrt{n}}\frac{\bar{Y}_{3-j}(s,\gamma^o)}{\bar{Y}_{\cdot}(s,\gamma^o))}\right\}^2 Y_{ij}(s)\tilde{h}_{ij}(s,\beta^o)\frac{d}{ds}\Lambda_{0j}^{(n)}(s).$$

By writing $\frac{d}{ds}\Lambda_{0j}^{(n)}(s)=\frac{d\Lambda_0(s)}{ds}\frac{d\Lambda_{0j}^{(n)}(s)}{d\Lambda_0(s)}$ and using Assumption 5 we get

$$\sum_{j=1}^{2}\left\{D_j^n(s,\gamma^o)\right\}^2\lambda_j^n(s)\longrightarrow_p W^2(s)\frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}_1(s)+\bar{y}_2(s)}\lambda_0(s),\qquad as\quad n\to\infty.$$

Now, by Assumptions 2, 5 and 6 we obtain that

$$\sup_{s\in[0,\tau]}\left|\sqrt{n}\Big\{\frac{\tilde{Y}_1(s,\gamma^o)\bar{Y}_2(s,\gamma^o)d\Lambda_{01}^n(s)}{\tilde{Y}_2(s,\gamma^o)\bar{Y}_1(s,\gamma^o)d\Lambda_{02}^n(s)}-1\Big\}-\varphi(s)\right|\to 0.$$

Therefore, the second term of the right-hand side of (4.48) converges to $\mu_1(t)$ in probability, as $n\to\infty$.

4.8 Simulation

In this section we present our simulation study aimed to investigate the finite sample properties of our proposed procedures. The simulations were carried out under the popular Gamma frailty model. Therefore, we start by presenting the above procedures under the Gamma distribution with mean 1 and variance θ.

The log-likelihood function (4.3), under the frailty model with $Gamma(\frac{1}{\theta},\frac{1}{\theta})$, becomes

$$\begin{aligned}\ln L(\gamma)\propto&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\delta_{ij}\beta^T Z_{ij}-\sum_{i=1}^{n}\left\{\frac{\ln(\theta)}{\theta}+\left(N_{i\cdot}(\tau)+\frac{1}{\theta}\right)\ln\left(H_{i\cdot}(\tau)+\frac{1}{\theta}\right)\right\}\\&+\sum_{i=1}^{n}\left\{\ln\prod_{j=0}^{N_{i\cdot}(\tau)-1}(j+\frac{1}{\theta})\right\}I_{\{N_{i\cdot}(\tau)\geq 1\}},\end{aligned}\tag{4.49}$$

and the conditional mean and variance are

$$\psi_i(t) = E\Big(w_i \mid \mathcal{F}_{t-}\Big) = \frac{N_{i\cdot}(t-) + \theta^{-1}}{H_{i\cdot}(t-) + \theta^{-1}},$$

and

$$Var\Big(w_i \mid \mathcal{F}_{t-}\Big) = \frac{N_{i\cdot}(t-) + \theta^{-1}}{\{\hat{H}_{i\cdot}(t-) + \theta^{-1}\}^2}.$$

In the Gamma frailty model the parameter θ quantifies the strength of the dependence between event times of the same customer. As θ becomes large, the strength of dependence increases. We consider three levels of dependence: independence ($\theta = 0.01005$), mild dependence ($\theta = 1$) and strong dependence ($\theta = 4$). These values of the frailty parameters were defined based on the Kendall's τ coefficient (Kendall [49]). Under the Gamma frailty distribution Kendall's τ equals $\theta/(\theta + 2)$. Therefore, the respective values of Kendall's τ are as follows: $1/200$, $1/3$ and $2/3$. We assume constant baseline hazard functions $\lambda_{01}(t) = \lambda_{02}(t) = 1$ $t \in [0, \infty)$ and $\beta = (1, 2)^T$. In the following, we provide a detailed description of the sampling design used in the simulation study for sampling 2 calls for n customers.

1. Generate independent realizations $Z_{ij} \sim Uniform\{1, 2, 3\}$ $i = 1, ..., n$, $j = 1, 2$.

2. Generate n independent realizations of w from $Gamma(\frac{1}{\theta}, \frac{1}{\theta})$.

3. Generate n independent pairs of survival times (T^0_{i1}, T^0_{i2}) such that

$$T^0_{ij} \mid Z_{ij}, w_i \sim Exponential\{w_i \exp(\beta^T Z^*_{ij})\}, \qquad i = 1, ..., n \quad j = 1, 2,$$

where $Z^*_{ij} = (Z^{(1)}_{ij}, Z^{(2)}_{ij})^T$ and for $k = 1, 2$

$$Z^{(k)}_{ij} = \begin{cases} 1, & \text{if} \quad Z_{ij} = k \\ 0, & \text{otherwise} \end{cases}$$

4. Generate independent censoring times $C_{ij} \sim Exponential(3)$ $i = 1, ..., n$ $j = 1, 2$. Such a design yields $70\% - 80\%$ censoring rate.

5. Evaluate the observed times (T_{i1}, T_{i2}) and the event status, δ_{ij}, as follows:

$$\begin{aligned} &\text{if} \quad T^0_{ij} \le C_{ij} \quad \text{then} \quad T_{ij} = T^0_{ij} \quad \text{and} \quad \delta_{ij} = 1 \\ &\text{if} \quad T^0_{ij} > C_{ij} \quad \text{then} \quad T_{ij} = C_{ij} \quad \text{and} \quad \delta_{ij} = 0. \end{aligned}$$

Table 4.1: Summary of parameter estimates $\{\hat{\theta}, \hat{\beta}, \hat{\Lambda}(t)\}$ based on 1000 simulated random datasets with $n = 250$ and 500.

	independence			mild dependence			strong dependence		
	true value	mean	SD	true value	mean	SD	true value	mean	SD
				n=250					
θ	0.01005	0.084	0.150	1	0.929	0.240	4	3.806	0.812
β_1	1	1.023	0.328	1	0.969	0.249	1	1.009	0.355
β_2	2	2.059	0.318	2	1.935	0.264	2	1.986	0.368
$\Lambda_{01}(t_1)$	0.005	0.005	0.003	0.005	0.005	0.003	0.005	0.005	0.003
$\Lambda_{01}(t_2)$	0.01	0.010	0.004	0.01	0.011	0.004	0.01	0.010	0.005
$\Lambda_{01}(t_3)$	0.05	0.049	0.016	0.05	0.053	0.015	0.05	0.050	0.018
$\Lambda_{01}(t_4)$	0.1	0.096	0.031	0.1	0.104	0.029	0.1	0.098	0.034
$\Lambda_{01}(t_1)$	0.005	0.005	0.003	0.005	0.006	0.003	0.005	0.005	0.003
$\Lambda_{01}(t_2)$	0.01	0.010	0.005	0.01	0.011	0.005	0.01	0.010	0.005
$\Lambda_{01}(t_3)$	0.05	0.049	0.017	0.05	0.054	0.015	0.05	0.051	0.017
$\Lambda_{01}(t_4)$	0.1	0.097	0.033	0.1	0.108	0.029	0.1	0.101	0.031
				n=500					
θ	0.01005	0.064	0.098	1	1.025	0.175	4	3.925	0.596
β_1	1	1.013	0.219	1	1.008	0.198	1	1.007	0.242
β_2	2	2.021	0.211	2	2.003	0.201	2	1.999	0.262
$\Lambda_{01}(t_1)$	0.005	0.005	0.002	0.005	0.005	0.002	0.005	0.005	0.002
$\Lambda_{01}(t_2)$	0.01	0.010	0.003	0.01	0.010	0.003	0.01	0.010	0.004
$\Lambda_{01}(t_3)$	0.05	0.049	0.011	0.05	0.050	0.010	0.05	0.050	0.013
$\Lambda_{01}(t_4)$	0.1	0.098	0.021	0.1	0.101	0.019	0.1	0.100	0.025
$\Lambda_{01}(t_1)$	0.005	0.005	0.002	0.005	0.005	0.002	0.005	0.005	0.002
$\Lambda_{01}(t_2)$	0.01	0.010	0.003	0.01	0.010	0.003	0.01	0.010	0.004
$\Lambda_{01}(t_3)$	0.05	0.049	0.011	0.05	0.050	0.010	0.05	0.051	0.014
$\Lambda_{01}(t_4)$	0.1	0.098	0.021	0.1	0.101	0.019	0.1	0.100	0.025

Table 4.2: *Comparison of* $\hat{\sigma}_I^2(t)$ *and* $\hat{\sigma}_{II}^2(t)$ *for* $n = 250$ *and* 500.

θ	minimum	1-st quartile	median	mean	3-rd quartile	maximum	
			n=250				
0.01005	$\hat{\sigma}_I^2(t)$	0.064	0.086	0.092	0.093	0.099	0.133
	$\hat{\sigma}_{II}^2(t)$	0.063	0.083	0.089	0.090	0.096	0.125
1	$\hat{\sigma}_I^2(t)$	0.077	0.118	0.130	0.131	0.142	0.234
	$\hat{\sigma}_{II}^2(t)$	0.072	0.108	0.117	0.118	0.128	0.190
4	$\hat{\sigma}_I^2(t)$	0.080	0.132	0.149	0.154	0.171	0.341
	$\hat{\sigma}_{II}^2(t)$	0.071	0.115	0.130	0.134	0.149	0.305
			n=500				
0.01005	$\hat{\sigma}_I^2(t)$	0.075	0.091	0.096	0.096	0.100	0.124
	$\hat{\sigma}_{II}^2(t)$	0.075	0.090	0.095	0.095	0.099	0.115
1	$\hat{\sigma}_I^2(t)$	0.107	0.129	0.138	0.138	0.146	0.180
	$\hat{\sigma}_{II}^2(t)$	0.098	0.117	0.124	0.124	0.131	0.162
4	$\hat{\sigma}_I^2(t)$	0.089	0.123	0.137	0.138	0.151	0.199
	$\hat{\sigma}_{II}^2(t)$	0.078	0.109	0.119	0.120	0.132	0.169

Finally we obtain $(T_{i1}, \delta_{i1}, Z_{i1}, T_{i2}, \delta_{i2}, Z_{i2})$ $i = 1, ..., n$. We consider $n = 250$ or 500 and $\tau = 0.1$. The results are based on 1000 random samples.

Tables 4.1 summarizes the results of $\{\hat{\theta}, \hat{\beta}, \hat{\Lambda}(t)\}$ and presents the true parameters' values, the empirical mean of the estimates and the standard deviation. For the cumulative baseline hazard functions $\hat{\Lambda}_{0j}(t)$ we consider the values at $t = 0.005, 0.01, 0.05$ and 0.1. Table 4.1 verifies that our estimating procedure performs well in terms of bias.

Table 4.2 compares the two variance estimators, $\hat{\sigma}_I^2(t)$ and $\hat{\sigma}_{II}^2(t)$, by presenting the following descriptive statistics: the minimum, 1-st quantile, median, mean, 3-rd quantile and the maximum. It is evident that the differences between the two estimators are very small even under a strong dependency such as $\theta = 4$. These results support our recommendation to use $\hat{\sigma}_{II}^2(t)$ rather than $\hat{\sigma}_I^2(t)$.

Now, we are interested in comparing between our proposed variance estimator of $S_n(t, \hat{\gamma})$ and other naive variance estimators. One is an estimator that does not take into account the dependence between the samples. We denote this estimator by $\hat{\sigma}_1^2(t)$ and it easy to verify that

$$\hat{\sigma}_1^2(t) = \frac{1}{n} \int_0^t \hat{W}_n(s) \sum_{j=1}^{2} \sum_{i=1}^{n} \left\{ \frac{\bar{Y}_{3-j}(s, \hat{\gamma})}{\bar{Y}_{\cdot}(s, \hat{\gamma})} \right\}^2 d\bar{N}_j(s).$$

Table 4.3: *Comparison of our proposed variance estimators with naive estimators*

		Naive		Song et al.		Proposed I		Proposed II	
	empirical SD		empirical		empirical		empirical		empirical
θ	of $S_n(t,\hat{\gamma})$	$\hat{\sigma}_1(t)$	Type I error	$\hat{\sigma}_2(t)$	Type I error	$\hat{\sigma}_I(t)$	Type I error	$\hat{\sigma}_{II}(t)$	Type I error
				n=250					
0.01005	0.292	0.297	0.045	0.298	0.045	0.304	0.038	0.299	0.040
1	0.335	0.312	0.064	0.312	0.066	0.361	0.030	0.343	0.037
4	0.330	0.280	0.098	0.279	0.100	0.390	0.013	0.364	0.024
				n=500					
0.01005	0.317	0.305	0.064	0.305	0.064	0.309	0.062	0.308	0.064
1	0.353	0.320	0.074	0.319	0.076	0.371	0.041	0.352	0.051
4	0.334	0.271	0.097	0.271	0.099	0.370	0.015	0.346	0.031

The second estimator is the robust estimator of Song et al. [79] and is given by

$$\hat{\sigma}_2^2(t) = \frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\Big\{\int_0^t \hat{W}_n(s)\mathcal{G}(s,\hat{\gamma})d\hat{M}_{ij}(s)\Big\}^2,$$

where $\hat{M}_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(s)e^{\hat{\beta}^T Z_{ij}}d\bar{N}_j(s)/\bar{Y}_j(s,\hat{\gamma})$ $i = 1, ..., n$, $j = 1, 2$. This estimator was proposed for repeated events where the two baseline hazard functions were estimated based on independent samples. In Table 4.3 we present the mean of each variance estimator and the empirical significance level of a test with Type I error $\alpha = 0.05$. The empirical significance level is the percent of tests such that the null was rejected. The results show that under the independent setting the four methods provide similar results, but as the dependence increases, the differences between our methods and the two other naive methods, tend to increase as well. The empirical significance level for the other estimators increases with θ. It is evident, that only our methods perform reasonably well under any dependency level. In some cases the empirical Type I error of the naive is about 9%. In addition, there are small differences between the empirical Type I error provided by our two proposed estimators $\hat{\sigma}_I^2$ and $\hat{\sigma}_{II}^2$. Hence our recommendation of using (4.22) for the variance estimate of $S_n(t,\hat{\gamma})$, is again being justified.

Now, we provide a simulation results to evaluate the proposed sample size formula. All the three levels of dependence were examined under a two-sided test, with $\alpha = 0.05$, and $\pi = 0.80$. The baseline hazard functions correspond to the local alternative (4.29), namely, $\lambda_{01} = \exp\{\varepsilon/2\sqrt{n}\}\lambda_0(s)$ and $\lambda_{02} = \exp\{-\varepsilon/2\sqrt{n}\}\lambda_0(s)$, where $\lambda_0(s) = 1$ and ε takes the values of 0.3, 0.5 and 0.6.

Table 4.4: *Empirical power of a two-sided test with* $\alpha = 0.05$ *and* $\pi = 0.80$.

	independence		mild dependence		strong dependence		
	sample	empirical	sample	empirical	sample	empirical	Schoenfeld's
ε	size	power	size	power	size	power	sample size
0.3	201	0.819	264	0.766	536	0.769	174
0.5	70	0.805	110	0.789	241	0.833	63
0.6	50	0.773	84	0.823	181	0.836	44

We first generated 100 random samples for each configuration, and based on these simulated data we calculated the average sample size based on (4.35). These results serve as the required sample size with $\alpha = 0.05$, and $\pi = 0.80$. Then, for each configuration we generated 1000 random samples with the respective sample size. For each sample we calculated the test statistic $S_n(t, \hat{\gamma})$ and its variance estimate $\hat{\sigma}^2_{II}(t)$. Finally, we calculated the empirical power based on a two-sided test with $\alpha = 0.05$, to be compared with the theoretical power of 0.80. The results are presented in Table 4.4.

Table 4.4 shows that our sample size formula performs well since the empirical power is reasonably close to the nominal power 0.80. The results demonstrate that as the difference between the two baseline hazard functions increases, less observations are required. The formula of Schoenfeld [76]: $(Z_{1-\alpha/2} + Z_\pi)^2/(2\varepsilon^2)$ gives similar values as our formula in the case of independence ($\theta = 0.01005$), as expected. In all other cases, Schoenfeld's formula under estimate the required sample size.

4.9 Data Analysis

In this section we present the analysis of customer patience based on data from a real call center. The data structure was explained in Section 4.1. The sample size of the data considered in the analysis is $49,246$ customers, with only one sequence of calls for each customer, and each customer had not called for at least two months before the beginning of the sequence. By this we hope to ensure that customers are not familiar with the current system at their first call. For each customer we consider up to 5 calls. Table 4.5 presents the distribution of the observed calls.

Table 4.5: *Summary of the call center data set.*

	1-st call	2-nd call	3-rd call	4-th call	5-th call
number of calls	49246	7759	1646	488	198
number of events	1416	360	89	32	18

Table 4.6: *The call center data set: parameters' estimates and bootstrap standard errors.*

	$\hat{\theta}$	$\hat{\beta}_1$	$\hat{\beta}_2$
point estimate	0.9973	-0.3006	-0.1211
bootstrap SE	0.1767	0.1046	0.1046

The covariate considered is a type of customer: 1 - for VIP, 2 - medium importance and 3 - standard customer. Hence, β_1 reflects the effect of VIP vs all others, and β_2 reflects the effect of medium importance vs others.

Table 4.6 presents the parameter estimates under the Gamma frailty model along with their bootstrap standard errors, based on 150 bootstrap samples. The results show that the frailty parameter is close to 1, meaning moderate dependence between calls of the same customer. The estimates of the regression coefficients indicate that if a customer is more important, then his/her chance to abandon before being served is lower than the chance of a less important customers.

In Table 4.7 we present the estimated values of the baseline hazard functions calculated at times: 10, 50, 100, 150, 200 and 250 seconds. According to the results, the values that belong to the first call are smaller than the values of the other calls, and the values that belong to the fifth call are larger than the other values. For visual inspection of the estimated baseline hazard functions the reader is referred to Figure 4.2.

Table 4.7: *The call center data set: Estimates of the cumulative baseline hazard functions.*

					exponential (I)					
	1-st call		2-nd call		3-rd call		4-th call		5-th call	
		bootstrap		bootstrap		bootstrap		bootstrap		bootstrap
t	$\hat{\Lambda}_{01}(t)$	SE	$\hat{\Lambda}_{02}(t)$		$\hat{\Lambda}_{03}(t)$	SE	$\hat{\Lambda}_{04}(t)$	SE	$\hat{\Lambda}_{05}(t)$	
10	0.012	0.001	0.010	0.002	0.009	0.003	0.009	0.005	0.030	0.012
50	0.027	0.002	0.039	0.004	0.034	0.007	0.026	0.009	0.057	0.022
100	0.051	0.003	0.085	0.007	0.080	0.015	0.065	0.016	0.151	0.048
150	0.075	0.004	0.152	0.014	0.132	0.023	0.155	0.046	0.178	0.062
200	0.108	0.006	0.221	0.020	0.174	0.029	0.266	0.069	0.305	0.105
250	0.148	0.009	0.301	0.026	0.256	0.040	0.407	0.107	0.553	0.183

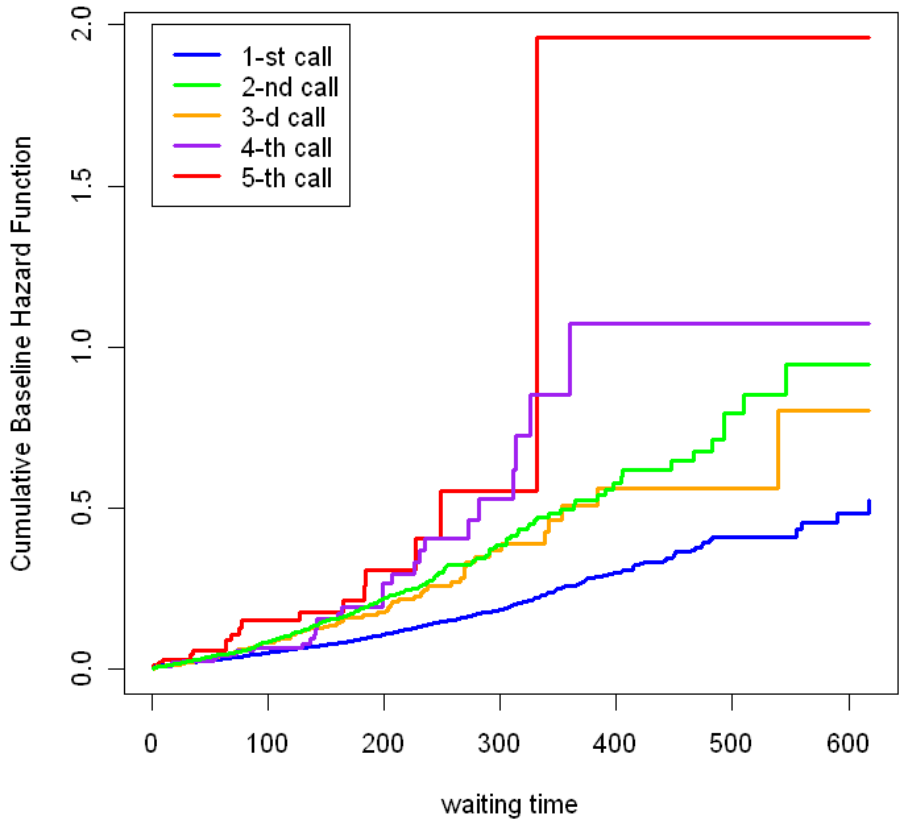


Figure 4.2: Estimates of the cumulative baseline hazard functions.

It is evident that the baseline hazard function of the first call is always below the other hazard functions, and the function of the fifth call is almost always above the others. For the other functions one could say that the differences are not so obvious.

Table 4.8: *The call center data set: results of the paired tests.*

calls	1 - 2	1 - 3	1-4	1-5	2 - 3
$S_n(250,\hat{\gamma})$	-0.464	-0.771	-0.051	-0.048	0.027
$\hat{\sigma}_{II}(250)$	0.039	0.199	0.018	0.016	0.027
$S_n(250,\hat{\gamma})/\hat{\sigma}_{II}(250)$	-11.915	-3.871	-2.884	-3.019	1.024
p-value	< 0.001	< 0.001	0.002	0.001	0.847
FDR p-value	< 0.001	< 0.001	0.042	0.003	1.000
calls	2 - 4	2 - 5	3-4	3-5	4 - 5
$S_n(250,\hat{\gamma})$	0.058	-0.029	-0.014	-0.030	-0.019
$\hat{\sigma}_{II}(250)$	0.163	0.016	0.016	0.015	0.012
$S_n(250,\hat{\gamma})/\hat{\sigma}_{II}(250)$	0.355	-1.841	-0.907	-2.051	-1.493
p-value	0.639	0.033	0.182	0.020	0.068
FDR p-value	1.000	0.096	1.000	0.042	0.422

Now we would like to answer the following question: "Are the functions presented in Figure 4.2 really different, or are all these functions merely different estimators of the same function?". To answer this question we apply our test for comparing between each two cumulative baseline hazard functions. The results of these test are presented in Table 4.8.

Table 4.8 we present the values of the test statistic $S_n(250,\hat{\gamma})$, the estimated standard error based on (4.22), the standardized test statistic, the p-value based on the standard normal distribution and the corrected p-value based on the FDR method [11] for correcting the dependent comparisons. The results show us that the baseline hazard function of the first call is significantly different from that of all other calls, even after correcting for multiple comparisons. There is also a significant difference between the baseline hazard functions of the third and the fifth calls. Differences between all the other functions are not statistically significant.

In the following we consider visual and naive way to compare two baseline hazard functions by using 95% pointwise confidence intervals. The interval for the baseline hazard function of each call j, $j = 1, ..., 5$, is created as follows: for each bootstrap sample we estimate the cumulative baseline hazard function. At each event time, we estimate the 0.025-quantile by the 4-th ordered estimate and the 0.975-quantile by the 146-th ordered estimate and these are our lower and upper bounds of 95% pointwise confidence interval.

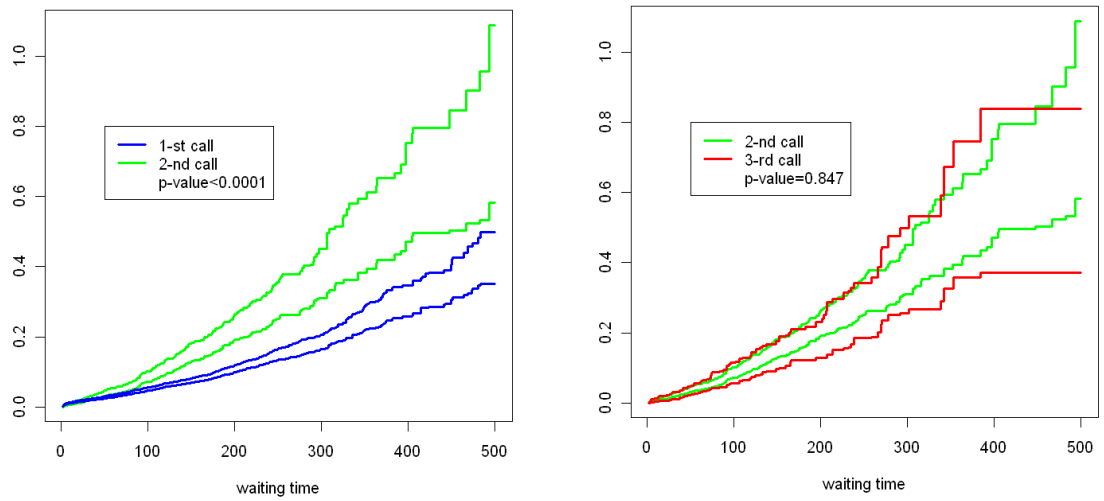


Figure 4.3: Naive 95% confidence intervals of the first and the second calls (left plot) and the second and the third calls (right plot).

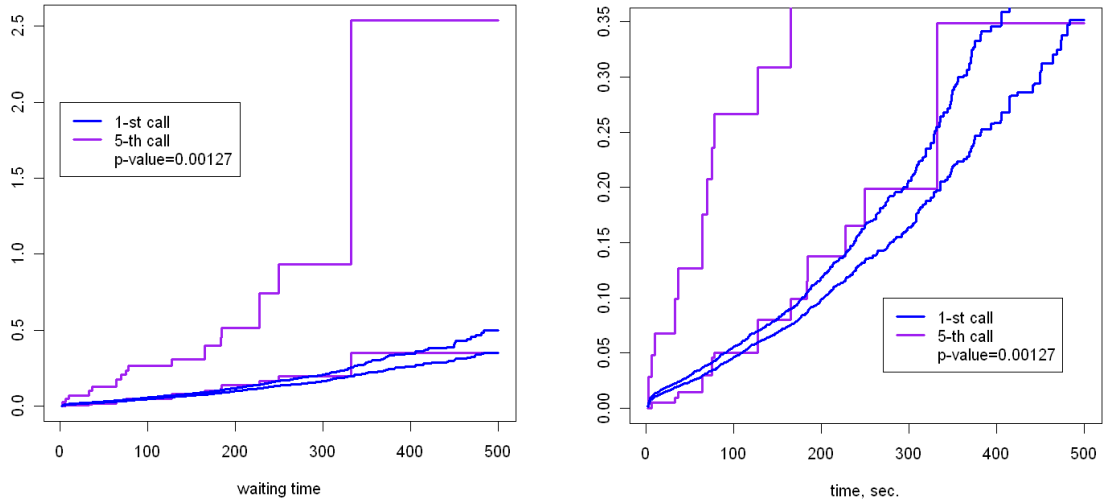


Figure 4.4: Naive 95% confidence intervals of the first and the fifth calls.

The left plot of Figure 4.3 presents the 95% confidence interval for the first (blue) and the second (green) calls. It is evident that there are no intersections between the two confidence intervals. The right plot of Figure 4.3 is for the second (green) and the third (red) calls, and Figure 4.4 is for the first (blue) and the fifth (violet) calls. The two plots of Figure 4.4 are similar but with a different resolution. It is clear that in such cases as of calls 2 and 3, the conclusion is obvious, and no statistical test is required. However, for a case such as of calls 1 and 5, our test results provide a clear important information that the two baseline hazard functions are different.

4.10 Summary and Future Directions

In this chapter, we considered a model for customer patience. The proposed model is an extended Cox model with frailty variate, reflecting the heterogeneity of customers, and different baseline hazard functions, reflecting the customer's familiarity with the system. For estimation of parameters, the method proposed by Gorfine et al. [37] was extended to the case of different baseline hazard functions. The simulation study indicated that our method works well in terms of bias for finite samples with any level of dependency within customer calls.

We provided a test for comparing two or more baseline hazard functions. The asymptotic distribution of the proposed test statistic was presented and a simulation study was conducted. The results of the simulation study show that our proposed method works well and as expected, gives better results in compare to the naive approaches that ignores within-subject dependence.

A sample-size formula was derived based on the limiting distribution of our test statistic under local alternatives. Our simulation study shows that according to the proposed formula the empirical power is reasonably close to the nominal value.

The proposed approach was applied to a real call center dataset and it was found that customers are significantly more patient in their first call. Moreover, customers that are defined as more important to the system, are willing to wait longer than the less important customers. In addition, there is a moderate level of dependence in the waiting behavior of a customer.

4.10.1 Application of the Proposed Approach in Health Care Data

This research project was motivated by the analysis of call center data, but our test can also be useful in other research areas. For example, consider the Washington Ashkenazi Kin-Cohort Study (WAS) (Struewing et al. [81]). In this study, blood samples and questionnaire were collected from Ashkenazi Jewish men and women volunteers living in the Washington DC area. Based on blood samples, volunteers were tested for specific mutations in BRCA1 and BRCA2 genes. The questionnaire included information on cancer and mortality history of the first-degree relatives of the volunteers.

For the current analysis we consider a subset of the data consist of female first-degree relatives of volunteers (mother, sisters and daughters). The *event* is

the age at diagnosis of breast cancer, and the covariate is the presence or absence of any BRCA1/BRCA2 mutations in the volunteer's blood sample. The data consist of $4,835$ families with $1-8$ relatives and a total of $13,030$ subjects.

So far, these data were analyzed under the assumption that the baseline hazard functions are identical to all family members: mother, daughters and sisters. We want to alow for each generation to have its own baseline hazard function, where the volunteer's generation is defined based on year of birth: before 1930 or otherwise.

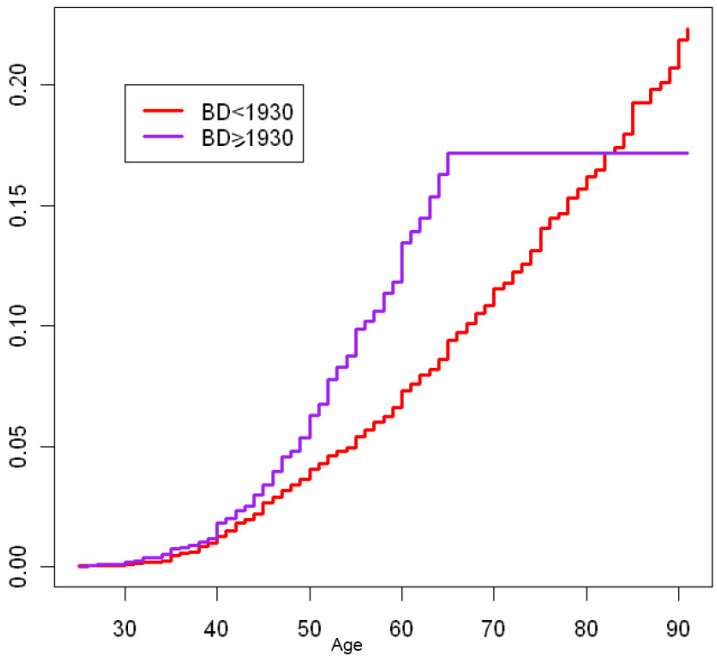


Figure 4.5: Estimates of the cumulative baseline hazard functions for the WAS data by birth year.

We start with reporting on the point estimates. The estimated frailty parameter under the $Gamma(\frac{1}{\theta},\frac{1}{\theta})$ model equals $\hat{\theta}=1.86$, the regression coefficient equals $\hat{\beta}=1.39$ and the estimates of the cumulative baseline hazard functions are presented in Figure 4.5. The estimated parameter of the frailty distribution indicates high dependence among family members. In the near future, we plan to estimate the standard errors of estimators. It is evident that the baseline risk of the older generation is always lower than that of the younger generation. Such a finding supports other publications reporting that cancer rates have risen in the past years [17], and it is likely that such a tendency is also as a result of

the increase in screening programs which detect the cancer in earlier stages [69]. As the continuation of this study we also plan to apply our statistical test for comparing the two hazard functions. We expect to verify our visual inspection and observe a significant difference between the two functions.

Another possible example of our model in a medical context is the scenario where patients suffer a series of events that require hospitalization, and we are interested in whether the distribution of the length of the k-th hospital stay depends on k. In this example, though, time probably will be discrete.

4.10.2 Future directions

Our estimation method and statistical test are not limited to a particular frailty distribution. The simulations and the data analysis were done under the Gamma frailty model distribution. It could be of importance to see how the results of the data analysis will vary, if at all, with other choices of frailty distribution. Also, it is important to check the effect of using a wrong frailty distribution. For example, the frailty is log-normally distributed, but the analysis is done under the Gamma distribution.

An important extension of our approach is the prediction of customer patience. An implementation of the prediction of customer patience in the modern Customer Relationship Management (CRM) software tools could be a huge advance in management of call centers. Such an option can significantly improve customers' satisfaction without additional financial costs. However, the right implementation of this feature is not a simple task and it could rise additional questions related to management science and queuing theory. Another possible extension of our proposed model could be including of time dependent covariates.

Another future direction in customer patience analysis is to analyze the changes in customer patience with the help of the hazard function. Adjusting the definition in [42] to our case, at any time point t, the hazard function is defined as the probability of abandonment within a short interval, given that the customer was in a queue at the beginning of the interval.

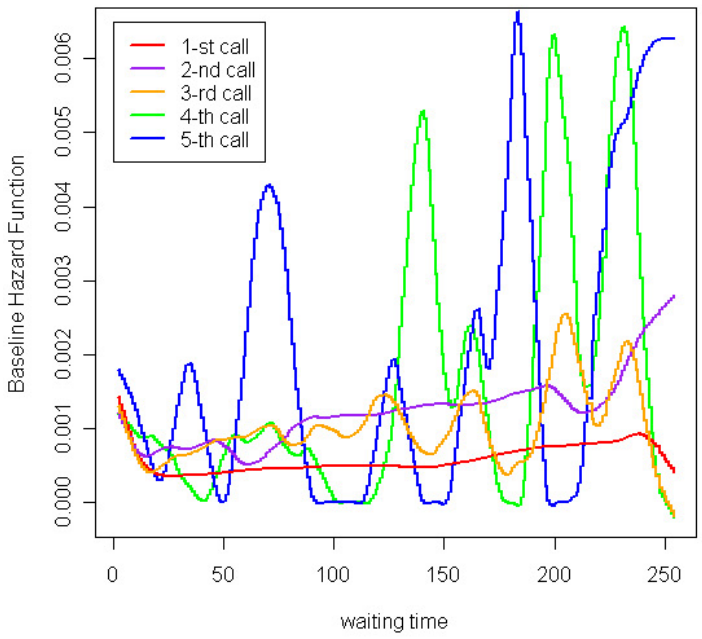


Figure 4.6: Estimates of the baseline hazard functions for the call center data.

Figure 4.6 presents smooth hazard functions for customer patience while waiting for agent service based on the analysis presented in Section 4. These functions are calculated as the derivatives of the smoothed cumulative baseline hazard functions presented in Section 4.9. The results indicate that customer patience distribution is not a monotone function, and it can be considered as a process developing over time. Figure 4.6 demonstrates that the baseline hazard function of the 1-st call almost linear and lies almost always under all the other baseline hazard functions. All the other functions are completely different and have a number of oscillations during the considered time period. More profound survival analysis of the behavior of such hazard functions will be a very interesting direction in the analysis of customer patience.

A possible disadvantage of our model (4.1) is the assumption that customer patience changes with the number of call consistently for all customers, while, it could be that these changes have individual features as well. Thus, we could extend our model and include two random factors: a frailty variate w_i at the customer level, and a frailty variate v_{ij} at a call level of each customer i ($i = 1, ..., n$). Such a model was considered by Aalen et al. [4]. It assumes that random effects operate multiplicatively on the baseline hazard, and conditional on the frailties w_i and v_j, the hazard function of customer patience at call j is of

the form

$$\lambda_{ij}(t|w_i, v_{ij}) = w_i v_{ij} \lambda_{0j}(t) e^{\beta^T Z_{ij}}, \quad j = 1, ..., m_i, \quad i = 1, ..., n, \tag{4.50}$$

where, for each customer, v_{ij} $j = 1, ..., m_i$ are i.i.d. random variables with density function $g(v) \equiv g(v; \mu)$ where μ is an unknown vector of parameters. The frailty v_{ij} can also be regarded as the set of covariates of call j that are not included in Z_{ij} because they are not measured. This is a hierarchical frailty model with two-levels of frailty: shared frailty at the customer level and unshared frailty at the call level of each customer.

We grouped the customer's calls in "series of retrials". Hence, for studying the effect of "series of retrials" we can consider the following model

$$\lambda_{ijk}(t|w_i, y_k) = w_i y_k h_{j0}(t) e^{\beta^T Z_{ijk}}, \quad k = 1, ..., l_i, \ j = 1, ..., m_{k_i}, \ i = 1, ..., n, \tag{4.51}$$

where y_k is the frailty variate of the k-th series of a specific customer, l_i is the total number of series of the customer, and m_k is the total number of calls of series k. This is also a hierarchical frailty model with two-levels of frailty: shared frailty at the customer level and shared frailty at the "series of retrials" level of each customer.

The hierarchial frailty model (4.51) is also considered by Aalen et al. [4] under frailty distributions determined by non-negative Lévy processes, which includes Power Variance Function (PVF) distributions. The PVF distributions include the gamma, positive stable, inverse Gaussian and compound Poisson distributions as special cases (see [1] and [42] for details). The implementation of such models can be a possible direction for further analysis.

The process of waiting on line before being served can be affected by factors developing with the time. For example, at the beginning a customer is expecting to wait a specific period of time, after this period s/he is astonished and after a while even angry of having to wait. Our frailty model is constant in time, but there may be a more realistic model which assumes a frailty that develops with time. This can be modeled by considering frailty as a stochastic process. Conditional on the unobserved frailty variate $W(t)$, the hazard function of customer patience would be of the form

$$h_0(t) W(t) e^{\beta^T W_j}, \qquad j = 1, ..., m_i. \tag{4.52}$$

Here, $W(t)$ is a stochastic process.

Chapter 5

DISCUSSION AND CONCLUSIONS

Call centers are intended to provide and improve customer service, marketing, technical support, etc. Therefore, the right management of a call center is a very important and crucial issue, that has to take into account many aspects. In our work we constructed and analyze an analytical model of a typical call center.

Operational performance measures, such as the probability for a busy signal, the probability of abandonment and average wait for an agent were calculated in this work. The calculations of these measures are cumbersome and they lack of insight. We thus approximated the measures in the QED asymptotic regime, which is suitable for moderate to large call centers. The approximations are easy to calculate for any number of agents.

A detailed comparison between exact and approximated performance shows that the approximations often work perfectly well, even outside the QED regime. Summarizing our findings through practical rules-of-thumb (expressed via the offered load). These rules were derived via extensive numerical analysis of our analytical results.

The approximations that have been developed can support the operations management of a call center, for example when trying to maintain a pre-determined level of service quality. We analyzed approximations of a real call center by models with and without an IVR, in order to evaluate the value of adding an IVR. Using real call center data we provided an analysis which was intended to connect theoretical investigations to real management problem solving and to allow the evaluation of the quality and robustness of the analytical results.

Our data analysis showed that the assumptions of exponentially distributed

service times does not take place, neither for IVR service nor for agent service. Similarly, the assumption that the arrival process is homogeneous Poisson is also overly simplistic. This problem was solved by division of the day into half-hour intervals. In this way, we find that within each interval the arrival rate is more or less constant and thus, within such intervals, we treated the arrivals as conforming to a Poisson process. The validation of our models against the US Bank Call Center demonstrated that the accuracy of the approximations is satisfactory. The approximated values in many intervals are very close to actual performance measures.

An extensive analysis of real call center data shows that a little change of parameters can dramatically change performance. Thus, the second part of our work is devoted to analysis of our model primitive, namely, customers patience, which is treated as a *process*. The provided study is a first attempt to apply frailty models to customer patience analysis. We suggested a novel statistical model and estimation procedure that was investigated theoretically and by extensive simulation studies. This model, together with the evident characteristics, allows taking into account personal customer features and customer experience with the system. This model provides an advance in customer patience analysis. We provided a computer program which enables convenient application and the possibility of processing large data samples by using our method of analysis.

We proposed a new test for comparison of two or more nonparametric baseline hazard functions considering dependent observations. Our test helps to analyze the influence of customer experience on his/her waiting behavior. The possible extension of our results may allow call center managers to define appropriate routing and priority rules for arriving calls on two different levels: for all customers and for each customer personally.

This thesis combines developing novel statistical models and tests, and analysis of real data sets. We expect that our research will contribute to a better understanding of customer habits, needs and expectations and this will have an impact on the improvement of call center operations, providing high quality service with lower costs. Therefore, we expect our work to be of both practical and theoretical importance.

We believe that some of the statistical models and procedures that were developed in this work, in particular the ones for customer patience analysis, can be applied in many other areas such us medicine, economics and industry where survival analysis is a very popular tool.

APPENDIX

A.1 The Estimation Procedure

```
#############################################################################
# Name: estimation.R
#
# Purpose: To find estimates of the our proposed model
#
# Arguments:
#          n:     number of customers;
#     J:     number of calls;
#     m:     number of observed times;
#          theta: frailty parameter;
#          beta:      vector of regression's coefficients;
#          lam:   matrix of cumulative baseline hazard functions estimates
#           with dimension (mx(J+1));
#          T:     matrix of observed times (nxJ);
#          z:     matrix (nxJ);
#          D:     matrix (mx(J+1));
#          delta: matrix of indicators of events (nxJ).
#############################################################################


source("functions.R")

x=c(theta,beta)
T<-T.data(data)
z<-Z.data(data)
D<-D.data(T,data)
delta<-delta.data(data)
lam<-lam.est(data,T,theta,delta,z,D,beta)
J=dim(T)[2]
a<-lam[,1]
```

```
est.lam<-matrix(rep(0,(length(a))*(J+1)),length(a),(J+1))
x=c(0.5,1,1)  # initialization of estimated parameters

  theta.est <- x[1]
  beta.est <- x[2:length(x)]
  repeat{
    est.theta.old<-x[1]
    repeat{

      beta.est.old<-x[2:length(x)]
      est.lam.old<-est.lam

      est.lam<-lam.est(y.data,T,theta.est,delta,z,D,beta.est)
            estb<-nlminb(x[2:length(x)],obj=lnLb,dat=data0,theta=x[1],
            z,lam=est.lam,delta,T=T)
      beta.est<-estb$par  #result of optimization procedure

      diff1<-max(abs(est.lam-est.lam.old)) # calculation of differences
      diff2<-max(abs(beta.est-beta.est.old))
            x[2:length(x)]<-beta.est
      if ((diff1 < 10^(-3)) & (diff2 < 10^(-5))) break
    }
   estt<-nlminb(x[1],obj=lnLt,dat=data0,beta=x[2:length(x)],z,lam=est.lam,delta,T)
    theta.est<-estt$par   # result of optimization procedure

diff3=abs(theta.est-est.theta.old)
    x[1]<-theta.est

    if (diff3<(10^(-5))) break
  }
est.lam<-lam.est(y.data,T,x[1],delta,z,D,x[2])

}
```

A.2 The Cumulative Baseline Hazard Functions

```
##############################################################################
# Name: functions.R
#
# Purpose: Main functions used in estimation procedure.
#
# New arguments:
#         tt:             vector of observed times;
#         lampred:        matrix of estimated values of cumulative baseline hazard functions,
#                         calculated at observed times, using as values at the previous step;
#         lampred_f:      matrix of estimated values of cumulative baseline hazard functions,
#                         calculated at observed times, using as values at the current step;
#         e:              matrix of exponents with power of product of regression's coefficients and
#                         covariets;
#         H:              matrix of values of function H defined in Section 4.4;
#         N:              matrix of number of events over each call of each customer;
#         psi:            vector of estimations of the frailty values;
#         lam:            matrix of cumulative baseline hazard functions estimates;
#         dlam:           matrix of jump values of cumulative baseline hazard functions;
#         L:              loglikelihood function;
##############################################################################

#########################################################################
#Calculation of the estimation for cumulative baseline hazard function
#########################################################################

lam.est<-function(y.data,T,theta,delta,z,d,beta){
##vector of event times
J=max(y.data[,5])
a<-y.data[y.data[,3]==1,]
a<-a[!duplicated(a[,2]),2]
a<-a[order(a)]
#############
##vector of all times
tt=T[,1]
for(j in 2:J){
tt<-c(tt,T[,j])
}
```

```
tt<-tt[!duplicated(tt)]
tt<-sort(tt)
if(tt[1]>0){tt=c(0,tt)}
last_t=tt[length(tt)]
n=dim(T)[1]

if(tt[1]>0){tt=c(0,tt)}
lampred<-matrix(rep(0,(length(tt))*(J+1)),length(tt),(J+1))
lampred_final<-matrix(rep(0,(n)*J),n,J)
lampred[1:length(tt),1]<-tt
q=1
qq=c(1,1)
if(a[1]>0){a=c(0,a)}
dlam<-vector()
lam<-matrix(rep(0,(length(a))*(J+1)),length(a),(J+1))
lam[,1]<-a
e<-e.data(z,beta,J)
N<-matrix(rep(0,n*J),n,J)
H<-matrix(rep(0,n*J),n,J)
b<-vector()
f<-vector()
y=NA
da1=matrix(rep(0,(J+1)*n),(J+1),n)
da2=matrix(rep(0,(J+1)*n),(J+1),n)
up<-vector()
down<-vector()
down1<-vector()
dlam1<-vector()

###### Update of values of the baseline hazard functions #######
for(i in 2:(length(a))){

repeat{
lampred[qq[1],2]=lam[i-1,2]
qq[1]=qq[1]+1
if(lampred[qq[1],1]>=a[i] | lampred[qq[1],1]>a[length(a)]
|lampred[qq[1],1]==max(T[,1])) break
}
```

```
###### Update of values of the baseline hazard functions #######
for(i in 2:(length(a))){
        repeat{
        lampred[qq[1],2]=lam[i-1,2]
        qq[1]=qq[1]+1
        if(lampred[qq[1],1]>=a[i] | lampred[qq[1],1]>a[length(a)]
        |lampred[qq[1],1]==max(T[,1])) break
        }
        repeat{
        lampred[qq[2],3]=lam[i-1,3]
        qq[2]=qq[2]+1
        if(lampred[qq[2],1]>=a[i] | lampred[qq[2],1]>a[length(a)]
        |lampred[qq[2],1]==max(T[,2])) break
        }
da1=lampred[match(T[,1],lampred[,1]),]
w1=da1[,1][!da1[,1] %in% y]
w2=da1[,2][!da1[,2] %in% y]
lampred_final[,1]=ifelse(w1<a[i-1],w2,lampred[lampred[,1]==a[i-1],2])
da2=lampred[match(T[,2],lampred[,1]),]
ww1=da2[,1][!da2[,1] %in% y]
ww2=da2[,3][!da2[,3] %in% y]
lampred_final[,2]=ifelse(ww1<a[i-1],ww2,lampred[lampred[,1]==a[i-1],3])
taub=rep(a[i-1],J)
down<-rep(0,J)
dlam<-rep(0,J)
############ calculation of denominator ##############
        N<-delta[,]*ifelse(T[,]<=taub,1,0)
    H[,]=lampred_final[,]*e[,]
    psi<-(rowSums(N)+(1/theta))/(rowSums(H)+(1/theta))
        for(j in 1:J)
        {
        up[j]<-(D[i-1,j+1])
        down[j]<-sum(psi*(ifelse(T[,j]>=a[i],1,0)*e[,j]))
        if(down[j]>0)dlam[j]=up[j]/down[j] else dlam[j]=0
        }
b=lam[i-1,2:(J+1)]
lam[i,2:(J+1)]<-b+dlam
}
 return(lam)
}
```

```
#############################################################################
#Calculation of the estimation for loglikelihood function
#############################################################################
lnLb<-function(beta,theta,z,lam,delta,T){
n=dim(T)[1]
J=dim(T)[2]
nnn=dim(lam)[1]
lnl=1
lam0=LM(lam[2:nnn,],T,J)
lam0=lam0[!lam0[,1]==0,]
e<-e.data(z,beta,J)
bt<-b.data(z,beta,J)
L<-0
H=0
LL<-vector()
LL=c(0)
a<-lam0[,1]
b<-a[length(a)]
#N<-N.c(delta,T,b)
N<-vector()
N<-rowSums(delta[,]*ifelse(T[,]<=b,1,0))
        for(i in 1:n){
H<-0
                for (j in 1:J){
                        if(T[i,j]>0){
                                if(delta[i,j]>0){
                                L<-L+delta[i,j]*bt[i,j]
                                }
                        H<-H+lam0[lam0[,1]==T[i,j],(j+1)]*e[i,j]
                        }
                }
        lnl=1
        for(m in 0:(N[i]-1))
                {
                lnl=lnl*(m+1/theta)
                }
        L<-L-(log(theta))/(theta)-(N[i]+1/theta)*log(H+1/theta)+
        ifelse(N[i]>0,log(lnl),0)
        LL=c(LL,L)
        }
return(-L)
}
```

A.3 Estimation for the Variance of $S_n\{\tau, \hat{\gamma}\}$

```
##############################################################################
# Name: functions.R
#
# Purpose: Main functions used in estimation procedure.
#
# New arguments:
#       S:          value of statitic;
#       V:          value of estimator of variance (naïve);
#       V1:         value of estimator of variance (Song et al.);
#       V2:         value of estimator of variance ( σ̂²_II(250) );
#       V3:         value of estimator of variance ( σ̂²_I(250) );
#       varw:       matrix of estimations of frailty calculated at each observable time;
#       dlam:       matrix of jump values of cumulative baseline hazard functions;
#       plam:       matrix of ψ̂_i(t) e^{β̂Z_ij} Y_ij(t);
#       blam:       matrix of e^{β̂Z_ij} Y_ij(t);
##############################################################################

source("functions.R")

a=D[,1]
a1=D[D[,2]==1,1]
a2=D[D[,3]==1,1]
DD=vector()
b=0
kkk=length(a[a[]<0.1])
k1=length(a1[a1[]<0.1])
k2=length(a2[a2[]<0.1])
item=matrix(rep(0,250*8),250,8)

dlam0<-ylam.data(data,T,theta,delta,z,D,beta)
dlam=dlam0
S0=sum((dlam[2:(kkk+1),3]*D[1:kkk,2]-
dlam[2:(kkk+1),2]*D[1:kkk,3])/(dlam[2:(kkk+1),3]+dlam[2:(kkk+1),2]))
for(k in 1:k1){
Y1=dlam[dlam[,1]==a1[k],2]
Y2=dlam[dlam[,1]==a1[k],3]
YY1=vector()
```

```
plam=plam.data(data,T,theta,delta,z,D,beta,a1[k])
blam=blam.data(data,T,theta,delta,z,D,beta,a1[k])

X1=ifelse(T[,1]<a1[k],0,Y2/(Y1+Y2)*(-plam[,1]/Y1))
X7=ifelse(T[,1]<=a1[k],0,((Y2/(Y1+Y2))^2)*plam[,1]/Y1)
X3=(Y2/(Y1+Y2))*(delta[,1]*ifelse(T[,1]==a1[k],1,0))*(1-plam[,1]/Y1)

item[,1]=item[,1]+ifelse(X3>0,X3,X1)
item[,3]=item[,3]+(Y2*Y1/(Y1+Y2))^2*ifelse(T[,1]<a1[k],0,
(blam[,1]/(Y1^3)))
item[,5]=item[,5]+X7
}

for(k in 1:k2){
Y1=dlam[dlam[,1]==a2[k],2]
Y2=dlam[dlam[,1]==a2[k],3]

plam=plam.data(data,T,theta,delta,z,D,beta,a2[k])
blam=blam.data(data,T,theta,delta,z,D,beta,a2[k])

X2=ifelse(T[,2]<a2[k],0,Y1/(Y1+Y2)*(-plam[,2]/Y2))
X8=ifelse(T[,2]<=a2[k],0,((Y1/(Y1+Y2))^2)*plam[,2]/Y2)
X4=(Y1/(Y1+Y2))*(delta[,2]*ifelse(T[,2]==a2[k],1,0))*(1-plam[,2]/Y2)

item[,2]=item[,2]+ifelse(X4>0,X4,X2)
item[,4]=item[,4]+(Y2*Y1/(Y1+Y2))^2*ifelse(T[,2]<a2[k],0,
(blam[,2]/(Y2^3)))
item[,6]=item[,6]+X8
}

x1=0
x2=0
x3=0
e<-e.data(z,beta,2)
a=D[,1]
b=0
varw=varw.data(data,T,theta,delta,z,d,beta,a[kkk])

varw1=varw
varw2=varw
```

```
for(s in 1:k){
        for(u in 1:k){
        varw=varw1[,max(u,s)]
        Y1=dlam[dlam[,1]==a[s],3]/((dlam[dlam[,1]==a[s],2]
          +dlam[dlam[,1]==a1[s],3])^2)

        Y2=dlam[dlam[,1]==a[u],3]/((dlam[dlam[,1]==a[u],2]
          +dlam[dlam[,1]==a1[u],3])^2)

item[,7]=ifelse(T[,1]>=max(a1[s],a1[u]),e[,1]*e[,1]*varw,0)
x1=x1+Y1*Y2*sum(item[,7])
                }
        }

for(s in 1:k){
        for(u in 1:k){
        varw=varw2[,max(s,u)]
        Y1=dlam[dlam[,1]==a [s],2]/((dlam[dlam[,1]==a2[s],2]
           +dlam[dlam[,1]==a2[s],3])^2)
        Y2=dlam[dlam[,1]==a2[u],2]/((dlam[dlam[,1]==a2[u],2]
           +dlam[dlam[,1]==a2[u],3])^2)

item[,7]=ifelse(T[,2]>=max(a2[s],a2[u]),e[,2]*e[,2]*varw,0)
x2=x2+Y1*Y2*sum(item[,7])
        }
}

for(s in 1:k){
        for(u in 1:k){
        if(u<s){varw=varw1[,s]}else{varw=varw2[,u]}
        Y11=dlam[dlam[,1]==a[s],3]/((dlam[dlam[,1]==a[s],2]
          +dlam[dlam[,1]==a[s],3])^2)
        Y21=dlam[dlam[,1]==a[u],2]/((dlam[dlam[,1]==a[u],2]
          +dlam[dlam[,1]==a[u],3])^2)

item[,8]=ifelse(T[,1]>=a1[s]&T[,2]>=a2[u],e[,1]*e[,1]*varw,0)
x3=x3+Y11*Y21*sum(item[,8])
        }
}
```

```
##
# Naive
##
V=sum(item[,5])+sum(item[,6])

##
# Song et al.
##
V1=sum(item[,1:2]^2)

##
# Our
##
V2=sum(item[,3])+sum(item[,4])

##
# Our full variance estimator
##
V3 =sum(item[,3])+sum(item[,4])+x1+x2-2*x3
```

A.4 Sample Size Calculation

```
##############################################################################
# Name: sample_size.R
#
# Purpose: Sample size calculation.
#
# New arguments:
#       T0:         expected value of statistic calculated according to the formula proposed in
#                   Section 4.6;
#       epsilon:    power of exponent reflecting the difference between the two baseline hazard
#                   functions;
#       Sigma2:     value of estimator of variance ( σ̂²_II(250) ).
##############################################################################

data <-read.table(dat[i],sep="",header=FALSE)
delta<-delta.data(data)
NN=100
lam0<-read.table(res[i],sep="",header=FALSE)
n=(dim(lam0)[1])/3

lam_0=cbind(lam0[1:n,1],lam0[(n+1):(2*n),1],lam0[(2*n+1):(3*n),1])
lam=lam_0[2:n,]
x=lam_0[1,]
theta<-x[1]
beta<-x[2:3]

T<-T.data(data)
z<-Z.data(data)
D<-D.data(T,dataw)
a=D[D[,1]<0.2,1]
kkk=length(a)
data=data_new
dlam0<-ylam.data(data,T,theta,delta,z,D,beta)
dlam=dlam0

T0=epsilon^2*((sum(delta[,1])+sum(delta[,2]))/NN)^2/4
TT=c(TT,T0)
a1=D[D[,1]<0.2&D[,2]==1,1]
a2=D[D[,1]<0.2&D[,3]==1,1]
k1=length(a1)
k2=length(a2)
```

```
item=matrix(rep(0,NN*8),NN,8)
for(k in 1:k1){
Y1=dlam[dlam[,1]==a1[k],2]
Y2=dlam[dlam[,1]==a1[k],3]


plam=plam.data(data,T,theta,delta,z,D,beta,a1[k])
blam=blam.data(data,T,theta,delta,z,D,beta,a1[k])

item[,3]=item[,3]+(Y2*Y1/(Y1+Y2))^2*ifelse(T[,1]<a1[k],0,
(blam[,1]/(Y1^3)))
}



for(k in 1:k2){
Y1=dlam[dlam[,1]==a2[k],2]
Y2=dlam[dlam[,1]==a2[k],3]
plam=plam.data(data,T,theta,delta,z,D,beta,a2[k])
blam=blam.data(data,T,theta,delta,z,D,beta,a2[k])

X4=(Y1/(Y1+Y2))*(delta[,2]*ifelse(T[,2]==a2[k],1,0))*(1-plam[,2]/Y2)

item[,4]=item[,4]+(Y2*Y1/(Y1+Y2))^2*ifelse(T[,2]<a2[k],0,(blam[,2]/(Y2^3)))
}
V2=c(V2,(sum(item[,3])+sum(item[,4]))/(NN))
Sigma2=(sum(item[,3])+sum(item[,4]))/(NN)
n_sample=c(n_sample,(1.9644854+0.84162)^2*sigma2/TT[i])
```

Bibliography

[1] Aalen, O.O., Borgan, O. and Gjessing, H.K. (2008). Survival and Event History Analysis. New York: Springer. 60, 76, 80, 98

[2] Aalen, O.O. (1994). Effects of Frailty in Survival Analysis. *Statistical Methods in Medical Research*, **3**, 227-243. 18

[3] Aalen, O.O. and Gjessing, H.K. (2001). Understanding the Shape of the Hazard Rate: A Process Point of View. *Statistical Science*, **16(1)**, 1-22. 18

[4] Aalen, O.O. and Moger, T.A. (2006). Hierarchical Lévy Frailty Models and a Frailty Analysis of Data on Infant Mortality in Norwegian Siblings. *UW Biostatistics Working Paper Series*, available at `www.bepress.com/cgi/viewcontent.cgi?article=1124&context=uwbiostat` 97, 98

[5] Acar, Crain and Yao (2010). Dependence Calibration in Conditional Copulas: A Nonparametric Approach. To appear in *Biometrics*, available at `http://www.utstat.utoronto.ca/craiu/Papers/R_DCCC_2.pdf` 68

[6] Aguir, M.S., Karaesmen, F., Akşin, O.Z., and Chauvet. F. (2004). The Impact of Retrials on Call Center Performance. *OR Spectrum, Special Issue on Call Center Management*, **26(3)**, 353-376. 15

[7] Andersen, P.K. and Gill, D.R. (1982). Cox's Regression Model for Counting Processes: a Large Sample Study. *Ann. Statist.*, **10**, 1100-1120. 65

[8] Andersen, P.K., Borgan, O., Gill, D.R. and Keiding, N. (1993). Statistical Models Based on Counting Processes. New York: Springer. 19, 70

[9] Armony, M., Gurvich, I. and Mandelbaum, A. (2008). Service Level Differentiation in Call Centers with Ffully Flexible Servers. *Management Science, Special Issue on Call Center Management*, **54(2)**, 279-294. available at

http://iew3.technion.ac.il/serveng/References/references.html
57

[10] Atar, R., Mandelbaum, A. and Shaikhet, G. (2006). Queueing Systems with Many Servers: Null Controllability in Heavy Traffic. *Annals of Applied Probability*, **16**, 1764-1804, available at
http://iew3.technion.ac.il/serveng/References/references.html
57

[11] Benjamini Y. and Yekutieli D. (2001). The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*, **29**, 1165-1188. 92

[12] Borst, S., Mandelbaum, A. and Reiman, M. (2004). Dimensioning Large Call Centers. *Operations Research*, 52(1), 17-34. 13

[13] Brandt, A., Brandt, M., Spahl, G. and Weber, D. (1997). Modelling and Optimization of Call Distribution Systems. *Proc. 15th Int. Teletraffic Cong.*, **15**, 133-144. 16, 25

[14] Bremaund, P. (1981). Point Processes and Queues. New York: Springer. 63

[15] Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, **53**, 1475-1484. 63

[16] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Zeltyn, S., Zhao, L. and Haipeng, S. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association*, **100**, 36-50. 17, 41, 42

[17] Center, M.,M., Jemal, A. and Ward, E. (2009). International Trends in Colorectal Cancer Incidence Rates. *Cancer Epidemiol. Biomarker Prev.*, 18(6), 1688-1694. 95

[18] Chen, H. and Yao, D.D. (2001). Fundamentals of Queueing Networks. New. York: Springer-Verlag. 27

[19] Chen, M. and Bandeen-Roche, K. (2005). A Diagnostic for Association in Bivariate Survival Models. *Lifetime Data Analysis*, **11**, 245-264. 18

[20] Cook, R.J., Lawless, J.F. and Nadeau, C. (1996). Robust Tests for Treatment Comparisons Based on Recurrent Event Responses. *Biometrics*, **52**, 557-571. 19

[21] Cox, D.R. (1972). Regression Models and Life Tables (with discussion). *J.R. Statist. Soc.*, **34**, 187-220, . 17, 60

[22] Donin, O., Trofimov, V., Feigin, P., Mandelbaum, A., Zeltyn, S., Ishay, E., Nadjarov, E. and Khudyakov, P. (2006). DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 4.1: The Call Center of "US Bank. Available at
`http://iew3.technion.ac.il/serveng/References/references.html`
12, 41

[23] Duffy, F.P. and Mercer R.A. (1978). A Study of Network Performance and Customer Behavior During Direct-Distance-Dialing Call Attempts in the U.S.A. *The BELL System Technical Journal*, **57**, 1-33, January 1978. 12

[24] Duffy, D.L., Martin, N.G. and Mathews, J.D. (1991). A Comparison of Clustered-Specific and Population Avareged Approaches for Analyzing Correlated Binary Data. *International Statistical Review*, **59**, 25-35. 18

[25] Duchateau, L. and Janssen, P. (2008). The Frailty Model. New York: Springer. 60

[26] Eng, K.H. and Kosorok, M.R. (2005). Sample size formula for the supremum log rank statistic. *Biometrics*, **61(1)**, 86-91. 19, 20

[27] Erlang, A.K. (1948). On the rational determination of the number of circuits. The life and works of A.K.Erlang. Copenhagen: The Copenhagen Telephone Company. 13

[28] Fleming, T.R. and Harrington, D.P. (1991). Counting Processes and Survival Analysis. New York: Wiley. 19, 20, 66

[29] Gangnon, R.E. and Kosorok, M.R. (2004). Sample-size formula for clustered survival data using weighted log-rank statistics. *Biometrika*, **91**, 2, 263-275. 19, 20, 72

[30] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing and Service Operations Management (MSOM)*, **5(2)**, 79-141. 13, 14, 17

[31] Garnett, O., Mandelbaum, A. and Reiman, M. (2002). Designing a Call Center with Impatient Customers. *Manufacturing and Service Operations Management (MSOM)*, **4(3)**, 208-227. 13, 14

[32] Gehan, E.A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrary Singly Censored Samples. *Biometrika*, **52**, 203-223. 19

[33] Gill, R.D. (1980). Censoring and Stochastic Integrals. Tract 124, Amsterdam: The Mathematical Center. 19

[34] Gilson, K.A. and Khandelwal, D.K. (2005). Getting more from call centers. The McKinsey Quarterly, Web exclusive, available at `http://www.mckinseyquarterly.com/article_page.aspx` 8

[35] Glidden, D.V. (1999). Checking the Adequacy of the Gamma Frailty Model for Multivariate Failure Times. *Biometrika*, **86**, 381-393. 18

[36] Glidden, D.V. and Vittinghoff, E. (2004). Modeling Clustered Survival Data form Multicenter Clinical trials. *Statistics in Medicine*, **23**, 369-388. 18

[37] Gorfine, M., Zucker, D.M. and Hsu, L. (2006). Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika*, **93**, 3, 735-741. 1, 62, 64, 65, 70, 94

[38] Halfin, S. and Whitt, W. (1981). Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, **29**, 567-587. 13

[39] Harris, C.M., Karla, L.H. and Saunders, P.B. (1987). Modeling the IRS Telephone Taxpayer Information System. *Operations Research*, **35(4)**, 504-523. 15

[40] Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *J.R. Statist. SOC. B*, **61**, 367-379. 18

[41] Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387-396. 18

[42] Hougaard, P. (2000). Analysis of Multivariate Survival Data. Springer-Verlag New York. 17, 18, 60, 96, 98

[43] Hsu, L., Chen, L., Gorfine, M. and Malone, K. (2004). Semiparametric Estimation of Marginal Hazard Function From the Case-Control Family Studies. *Biometrics*, **60**, 936-944. 18

[44] Hsu, L. and Gorfine, M. (2006). Multivariate Survival Analysis for Case-Control Family Data. *Biostatistics*, **7**, 387-398. 18

[45] Hsu, L., Gorfine, M. and Malone, K. (2007). On Robustness of Marginal Regression Coefficient Estimates and Hazard Functions in Multivariate Survival Analysis of Family Data when the Frailty Distribution is Misspecified. *Statistics in Medicine*, **26**, 4657-4678. 18, 19

[46] Jagerman, D.L. (1974). Some properties of the Erlang loss function. *Bell Systems Technical Journal*, **53(3)**, 525-551. 13

[47] Jelenkovic, P., Mandelbaum A. and Momcilovic P. (2004). Heavy Traffic Limits for Queues with Many Deterministic Servers. *QUESTA*, **47**, 53-69. 13

[48] Kalbfleisch, J.D. and Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data. Wiley New York. 17

[49] Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81-93. 85

[50] Khudyakov, P. (2006). Designing a Call Center with an IVR (Interactive Voice Response). M.Sc. Thesis, Technion, available at `http://iew3.technion.ac.il/serveng/References/references.html` 16, 27, 32, 34, 35

[51] Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algoritm. *Biometrics*, **48**, 795-806. 18

[52] Kocaga Y.L. and Ward A.R. (2009). Dynamic Outsourcing for Call Centers. *First Submitted January*, available at `http://www-rcf.usc.edu/~amyward/KW_Dynamic_Outsourcing.pdf` 37

[53] Kosorok, M.R., Lee, B.L. and Fine, J.P. (2004). Robust Inference for Univariate Proportional Hazards Frailty Regression Models. *The Annals of Statistics*, **32(4)**, 1448-1491. 18

[54] Kosorok, M.R. and Lin, C.Y. (1999). The Versality of Function-Indexed Weighted Log-Rank Statistics. *Jornal of the American Statistical Assosiation* **94**, 320-332. 20, 72

[55] Kort, B.W. (1983). Models and Methods for Evaluating Customer Acceptance of Telephone Connections. *IEEE*, 706-714. 12, 16

[56] Lawless, J.F. and Nadeau, C. (1995). Some Simple Robust Methods for the Analysis of Recurrect Events. *Technometrics*, **37**, 158-168. 19

[57] Liberman, P., Trofimov, V. and Mandelbaum, A. (2008). DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 5.1: Skills-Based-Routing-USBank. Available at `http://iew3.technion.ac.il/serveng/References/references.html` 45

[58] Lin, D.Y. (1994). Cox Regression Analysis of Multivariate Failure Time Data - the Marginal Approach. (1994). *Statistics in Medicine*, **13**, 2233-2247. 18

[59] Liu, K.S. (1980). Direct Distance Dialing: Call Completion and Customer Retrial Behavior", The BELL System Technical Journal, Vol.59, pp.295-311, March . 12

[60] Maman, S. (2009) Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments. M.Sc. Thesis, Technion, available at `http://iew3.technion.ac.il/serveng/References/references.html` 42

[61] Mandelbaum, A. (2008). Lecture Notes on QED Queues. Available at `http://iew3.technion.ac.il/serveng/Lectures/QED_lecture_Introduction_2008S.pdf` 35

[62] Mandelbaum, A., Massey, W.A., Reiman, M.I., Rider, B. and Stolyar, A.L. (2000). Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials. *Proceedings of the Fifth INFORMS Telecommunications Conference*, available at: `http://iew3.technion.ac.il/serveng/References/references.html` 15

[63] Mandelbaum, A. and Zeltyn, S. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems*, **51**, 361-402, available at
`http://iew3.technion.ac.il/serveng/References/references.html`
13, 44

[64] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, **50**, 163-170. 19

[65] Massey, A.W. and Wallace, B.R. (2006). An Optimal Design of the M/M/C/K Queue for Call Centers. To appear in *Queueing Systems*. 13, 14, 15, 27

[66] Murphy, S.A. (1994). Consistency in a Proportional Hazards Model Incorporating a Random Effect. *Ann. Statist.*, **23**, 182-98. 17, 18

[67] Nielsen, G.G., Gill, R.D., Andersen, P.K. and Sorensen, T.I. (1992). A counting process approach to maximum lekilihood estimation of frailty models. *Scand. J. Statist.*, **19**, 25-43. 18

[68] Palm, C. (1953). Methods of Judging the Annoyance Caused by Congestion. *Tele*, **2**, available at:
`http://iew3.technion.ac.il/serveng/Lectures/OptionalReading$_$8.html` 16

[69] Paltiel, O., Friedlander, Y., Deutsch, L., Yanetz, R., Calderon-Margalit, R., Tiram, E., Hochner, H., Barchana, M., Harlap, S. and Manor, O. (2007). The interval between cancer diagnosis among mothers and offspring in a population-based cohort. *Familiar Cancer*, **6**, 121-129. 96

[70] Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.*, **26**, 183-214. 17, 19, 63

[71] Peto, R. and Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society*, **135 (2)**, 185207. 19

[72] Prentice, R.L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167-179. 19

[73] Randhawa, R.S. and Kumar, S. (2009). Multi-Server Loss Systems with Subscribers. *Mathematics of Operations Research*, **34(1)**, 142-179. 16

[74] Ripatti, S. and Palmgren, J. (1978). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **65**, 153-158. 17, 18

[75] Roberts, J.W. (1979). Recent observations of subscriber behaviour. *9-th International Teletraffic Conference*. 12

[76] Schoenfeld, D.A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* **39**, 499-503. 20, 75, 89

[77] Shih, J.H. (1998). A Goodness-of-Fit Test for Association in a Bivariate Survival Model. *Biometrics*, **85**, 189-200. 18

[78] Shih, J.H. and Chatterjee, N. (2002). Analysis of Survival Data from Case-Control Family Studies. *Biometrics*, **58**, 502-509. 18

[79] Song, R., Kosorok, M.R. and Jianwen, C. (2008). Robust Covariate-Adjusted Log-Rank Statistics and Corresponding Sample Size Formula for Recurrent Events Data. *Biometrics*, **64**, 741-750. 19, 20, 74, 88

[80] Srinivasan, R., Talim, J. and Wang, J. (2004). Performance Analysis of a Call Center with Interacting Voice Response Units. *TOP*, **12**, 91-110. 10, 16, 26

[81] Struewing, J.P., Hartge, P., Wacholder, S., Baker, S.M., Berlin, M., McAdams, M, Timmerman, M.M., Brody, L.C. and Tucker, M.A. (1997).“The Risk of Cancer Associated with Specific Mutations of BRCA1 and BRCA2 among Ashkenazi Jews”. N *Engl J Med*, **336**, 1401-1408. 94

[82] Tarone, R.E. and Ware, J. (1977). On distribution-free tests for equality of survival distribution. *Biometrika*, **64**, 156-160. 19

[83] Trofimov, V., Feigin, P., Mandelbaum, A., Ishay, E. and Nadjarov, E. (2006). DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 1: Model Description and Introduction to User Interface. Available at `http://iew3.technion.ac.il/serveng/References/references.html` 12, 36, 41

[84] de Véricourt, F. and Jennings, O.B. (2008). Large-Scale Membership Services. *Operations Research*, **56**, 174-187. 15

[85] Weerasinghe, A. and Mandelbaum, A. (2009). Abandonment vs. Blocking in Many-Server Queues: Asymptotic Optimality in the QED Regime. Preprint. 38

[86] Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Mgmt. Sc.*, **38**, 708-723. 13

[87] Wolff, R.L. (1982). Poisson Arrivals See Time Averages. *Oper. Res.*, **30**, 223-231. 22

[88] Zeger, S., Liang, K.Y. and Albert, P.,S. (1998). Models for Longitudinal Data: A generalized Estimation Equation Approach. *Biometrics*, **44**, 1049-1060. 18

[89] Zeltyn, S. (2006). Call Centers with Impatient Customers: Exact Analysis and Many-Server Asymptotics of the M/M/n+G. Ph.D. thesis, Technion, available at
`http://iew3.technion.ac.il/serveng/References/references.html`
56

[90] Zeng, D. and Lin, D. Y. (2007). Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data (with discussion). *J.R. Statist. Soc. B*, **69(4)**, 507-564. 19, 64

[91] Zucker, D.M., Gorfine M. and Hsu L. (2007). Pseudo-full likelihood estimation for prospective survival analysis with a general semiparametric shared frailty model: Asymptotic theory. *J. Statist. Plann. Inference*, doi: 10.1016/j.jspi.2007.08.005.

19, 65