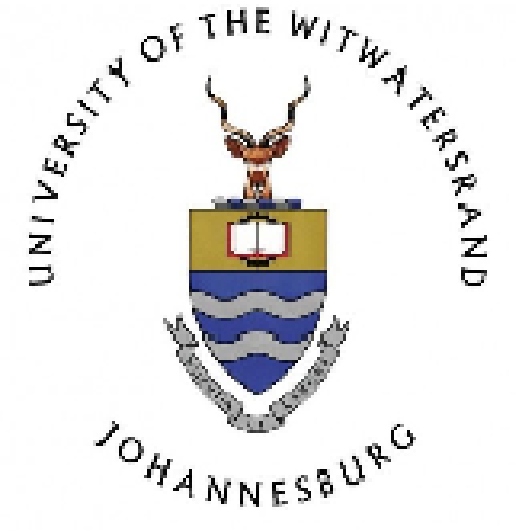


Optimization of Call Center Operations

BY

Lillian Kwesiga Kalenzi
Student Number: 308146

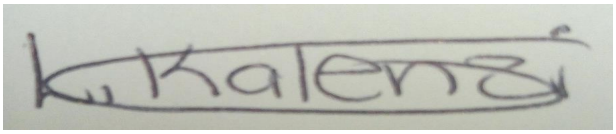
Supervisor: Dr HW Chipoyera

A research report submitted in partial fulfillment of the requirements for the degree of
Master of Science in Mathematical Statistics

SCHOOL OF STATISTICS AND ACTUARIAL SCIENCE
Faculty of Science
University of Witwatersrand
November 5, 2019

Declaration

I declare that this research work is my own, unaided work. It is being submitted for the degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at the University of the Witwatersrand or any other University.



SIGNATURE

DATE: November 2019

Abstract

In this work, an investigation into the problem of optimising the operations of a call center is done. The call arrival process is explored and found to be a non-homogeneous Poisson process with arrival rate that is a piece-wise constant function. The call service times are found to possibly be lognormally distributed.

The use of well-known queuing models such as the Erlang A, B and C in modeling a call center's operations with the ultimate goal of determining optimal number of agents needed to obtain an agreed upon targeted service level (SL) threshold is discussed. The target SL involves answering between 85% - 90% of all incoming calls within 15 /30 seconds as per industry norms.

Dedication

I dedicate this work to my loving parents and family.

Contents

List of Tables

List of Figures

Acknowledgments

The author wishes to express her deep and profound gratitude to:

My parents: in particular my father, Davidson Akiiki Kalenzi for all the never ending support and encouragement.

My family: Amani Tumusiime and dad, Siblings Sam, Emma and Kuse. Aunt Grace Onyango and family, Aunty Margret Birungi and family, Aunt Josephine Chipulu Mwewa and family, Grandmother Abwooli and family, Grandfather Joseph Musiitwa and family all cousin brothers and sisters and all other family members for all the prayers, love, support, patience and best wishes during this crucial and critical time.

My friends and colleagues, I thank you all for understanding and supporting me. I thank you for allowing me time to complete this work and for all the best wishes and words of encouragement during this time.

Special Thanks goes to my supervisor Dr HW Chipoyera for accepting and allowing me to research this work under his supervision and providing me with the guidance, encouragement and support to embark on this journey.

To all of you, thank you so much for believing in me and may the good Lord bless you all. "A Luta Continua" - the struggle continues, Victory is certain.

Chapter 1

Introduction

1.1 Introduction

Over the last few years, modeling of call center activities has attained a great deal of popularity. Nowadays, a telephonic call center is considered as an integral part of most organisations (Avramidis et al. (2004). Its main role is to maintain the operations of a company (or companies) so as to optimise its returns. It is utilised as a communication channel between the company and its clients. The call center is thus a bridge between the customers and the organisations for which the customers make their inquiries, voice their complaints and any market related issues.

Definition 1 *A* **telephone call center** *is a centralized or localized office consisting of a group of agents or personnel whose principal purpose is handling, routing, receiving and transmitting a high volume of customer calls related to their inquiries, queries, etc.*

Agents are usually qualified or trained employees who specifically handle the incoming calls of a call center. In other words a telephonic call center provides deliveries or solutions to customer queries through a network.

Most large organisations make use of telephonic call centers to sell products or render services to customers resulting in the industry maximising its profits and/or minimising its costs. There is therefore a high demand for call centers in organisations. A call center can either be physically located at the work-station (or offices) of a company (such an arrangement is known

as in-sourcing) or can be situated at a different place such as an individual's home or a geographically dispersed environment. The latter is known as a virtual call center (and is viewed as out-sourcing).

Call centers usually have two forms: *inbound* call center (handling incoming calls only) and *outbound* call center (handling outgoing calls only) or both, inbound and outbound.

1.2 Statement of The Problem

The problems explored in this research report are more or less the same problems bedeviling any call center. The main preoccupation of any call center manager is the desire to achieve a certain level of service.

In any business, customer perceptions play a critical role on the success of the business and have a huge impact on the returns of the business. Consequently the goal of every call center is to provide a desirable level of quality service to their customers. The downside of providing a quality level of service is that this comes at a cost, which if unchecked, can ultimately lead to business ruin. It is thus imperative that a quality level of service which takes into account a compromise between the cost of service to customers and retention of business in terms of returns is established. In order to achieve this, one needs to be cognizant of the issues that a call center manager grapples with which include:

- Trying to avoid the incidence of an unstable queuing system. A queuing system is said to be unstable if the number of calls in the system waiting for service (i.e. calls waiting to go through) grows indefinitely over time.
- The problem of an unstable queuing system is another challenge that the call center manager has to grapple with - that of some customers getting impatient and consequently dropping their calls. Call abandonment is a notorious feature of call center industries and the incidence of call abandonment is certainly a proxy of how bad a call centre's service is.
- Uncertainty about the number of CSRs (Customer Service Representatives or agents in other words) to deploy in order to achieve a good

level of service. Deploying a small number of CSRs may entail a shoddy service level whilst the deployment of too many CSRs may lead to the achievement of a desired level of service, and this may in turn result in the cost of providing the service ballooning to unacceptable levels.

1.3 Aim and Objectives of the Study

The aim of this research is to address the concern of providing a good quality level of queuing service at an inbound call center. To do so, one would need to have a good knowledge of some critical aspects, called operating characteristics of the queuing system such as

- the arrival pattern of the customers telephone calls; i.e. knowledge of the probability distribution of the number of incoming telephone calls per unit of time
- the queue service policy [1]
- the service rates of the CSRs, i.e. on average how many calls each server is capable of servicing per unit of time.
- the average amount of time a server takes to service a call (usually denoted by μ)
- the average number of calls waiting to be serviced, usually denoted by L_Q
- the average number of calls in the system, usually denoted by L
- the average amount of time a caller waits before they start being serviced, usually denoted by W_Q
- the average amount of time a caller spends in the system, usually denoted by W

[1] Some of the most common queue service policies include 1) First In First Out (abbreviated FIFO and probably the most common queue service policy - the first arriving call is processed first, 2) Last In First Out (LIFO) - last call arriving is processed first, 3) Service in random order (SIRO) - arriving calls are processed randomly, etc.)

- the probability distribution of the number of calls in the system; P_n is usually used to denote the probability that there are n calls in the system
- Grade of Service (see Definition 2)

Definition 2 (Grade of service) *The grade of service of a queuing system is the probability that all servers are busy when a call attempt is made.*

Note that, the Service Level Agreement (SLA) stipulates that the probability that an arriving call is answered within $15 - 30$ seconds is at least 0.85, i.e. at least 85% of calls received at a call center must be answered within 30 seconds. The specific objectives necessary to realize the aim of this research are

- to fit a probability distribution to the number of calls received per unit of time
- to determine the service rate of the CSRs, i.e. how many calls, on average, each of the servers successfully processes per unit of time
- to determine a queuing model which best describes the queue service facility at the call centre.
- to determine the probability distribution of call waiting times
- to determine the optimum number of agents required in order to satisfy the SLA.

1.4 Relevance of the Study

Koole and Mandelbaum (2002) mention that call centers are the preferred and most prevalent mode that companies use to communicate with their customers. They allude to the fact that the call center industry is vast and in terms of economic scope and workforce, it is rapidly expanding. They give the example of the United States and United Kingdom where estimates put 3% of the latter two countries' workforce as being involved with call centers. They also mention further that the call center industry realises an annual growth rate of 20%.

The modern day call center in South Africa has become indispensable in that it is the link between customers and South African companies/organisations (providing a vital service needed by the customers). As mentioned earlier on, the companies have an insatiable desire to achieve maximum levels of profitability or to offer a service at minimal cost. To maximise the companies' levels of profitability, call center managers are tasked with the responsibility of fine tuning their operating characteristics. Doing so partly entails having an adequate number of well trained agents and well serviced systems.

Angus (2001) gives a concise discussion of the need to neither deploy too few agents nor too many agents at any given time during operation. He argues that providing just one agent, while resulting in low cost in terms of *agent fees*, may lead to calls bunching up and queue instability and hence unacceptably poor service levels. On the other hand, having a large number of servers deployed, while it results in reduced call waiting times, will not make business sense because of the huge agents fees the call center would be saddled with.

In order to ensure that the clients do not hold or wait too long before being assisted, the call center manager needs to know the number of agents (who are well trained) that are needed in the call center. This helps the call center to be able to comply with the SLA which stipulates that at least 85% all incoming calls are answered within 15 - 30 seconds.

The downside to a call center being inefficient is that customers may end up highly frustrated and turning elsewhere for services. This would definitely have a negative impact on the companies' reputation and profitability levels.

Chapter 2

Literature review

Literature on statistical models for queuing systems in general is abundant; there is also a great deal of literature on call centers modeled as queuing systems. The classification and description of queuing models is done with the aid of Kendall's notation. Originally, the notation used three factors written in the form $A/S/c$ where A denotes the probability distribution of inter-arrival times for customers joining the queue, S the probability distribution of the time each server takes with a customer and c is the number of parallel servers. Further extensions have seen Kendall's notation transformed to $A/S/c/K/N/\mathcal{D}$ where K stands for the capacity of the queue, N is the size of the calling population and $\mathcal{D}$ depicts the queue discipline in force .

2.1 Some Common Queuing Systems

2.1.1 The $M/M/1$ Queuing Model

Usually a queuing system is classified as being either a Markovian or Non-Markovian system. A Markovian queuing model has exponentially distributed inter-arrival times and exponentially distributed service times. Results for Markovian models are considered under two categories: 1) transient state and 2) steady state. If probability distributions are time dependent, then they are said to be in transient state, otherwise they will be in a steady state.

The most elementary queuing system is described as the $M/M/1$ or $M/M/1/\infty/\infty/FIFO$ queuing model. In this model, arriving customers

queue up for service which is offered by a single server. The assumptions of the model are

- there is an infinite calling population whose members come to the facility with customers independently arriving at the facility in such a way that their arrival is not influenced by the queuing system;
- the number of arriving customers per unit of time follows a Poisson distribution with mean rate λ per unit of time; thus, the inter-arrival time of customers follows an exponential distribution with parameter $1/\lambda$;
- the queue is configured such that there is a single waiting line with unlimited space;
- the FIFO queue discipline is observed;
- the time the server takes to serve a customer follows an exponential distribution with mean $1/\mu$.

2.1.2 Other Commonly encountered Queuing Models

Other commonly encountered queuing models which assume the FIFO queue discipline and a queue configured in such a way that all arriving customers wait in a single line are given in Table 2.1.

Table 2.1: Other commonly encountered queuing models

	Queuing Model			
Characteristic	$M/M/n$	$M/M/n/n$	$M/G/n$	$M/G/n/n$
Calling Population	Infinite	Infinite	Infinite	Infinite
Arrival Process	Poisson	Poisson	Poisson	Poisson
Service time distribution	Exponential	Exponential	Non-exponential*	Non-exponential

* e.g. lognormal distribution

The $M/G/n/n$ model is a queuing model (with a provision for blocking) where arrivals are Markovian, service times have a *general distribution* and there are n servers; the Erlang B model gives a full account of this model.

2.2 Representation of a call center

Comprehensive literature surveys of call center modeling have been done by Brown et al. (2005). Figure 2.1[1] schematically represents a call center as a queuing system; the acronym ACD stands for Automated Calls Distributor.

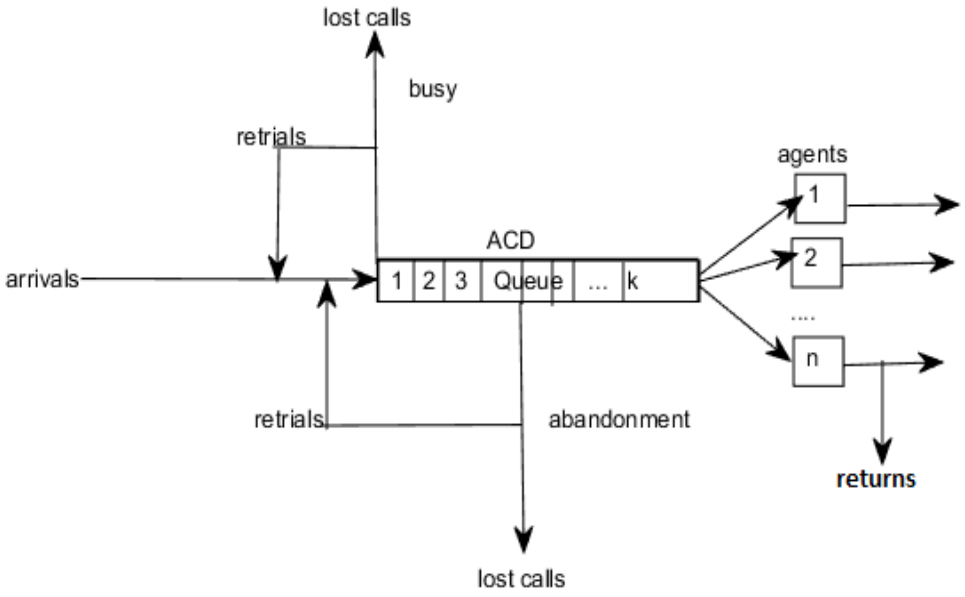


Figure 2.1: Schematic Diagram of a Call Center

2.3 History of mathematical formulation of call center problems

Angus (2001) asserts that Agner Krarup Erlang pioneered research on all modern day methods of optimizing the way networks operate while working for the Copenhagen Telephone Company in Denmark during the early part of the twentieth century. In his works, Erlang explores a queuing model whose

[1] originally presented in Mandelbaum 2000

arrival stream of calls is a Poisson process with rate λ and the service times follow an exponential distribution with parameter μ. In addition, there are n CSRs or agents who independently provide an identical service.

2.4 Erlang models

From literature, there are basically three popular Erlang queuing models: 1) the Erlang A Model, 2) the Erlang B Model, 3) the Erlang C Model; more recently, the Erlang R Model and the Hyper-Erlang models have been developed. According to Robbins et al. (2010), the most commonly used model by researchers and practitioners is the Erlang C model because of its simplicity but its drawback is that it ignores caller abandonment and a slight violation of any of its assumptions may lead to serious inaccuracies. The Erlang A model on the other hand allows for call abandonment; its major drawback is that performance measures associated with it are more difficult to calculate because they do not have closed form expressions.

2.4.1 Erlang C Model

For some time the Erlang C was the most popular Erlang model mainly because of its simplicity in that the results are tractable with closed form expressions. However, the Erlang C model has suffered criticism for making many assumptions which are deemed questionable in the context of a call center environment and has seen the Erlang A model overtaking it in terms of popularity. The major criticisms stem from the assumptions which are 1) the mean arrival rate being known, 2) there is no call abandonment and 3) service times follow an exponential distribution. According to Koole and Mandelbaum (2002), "$M/M/n$ predictions could turn out highly inaccurate because reality often "violates" its underlying assumptions, and these violations are not straightforward to model".

Remark 1 *The fundamental difference between the Erlang C and Erlang A models is that, with the Erlang C model, it is assumed that callers wait as long as it takes for them to receive service and do not abandon the queue whilst with the Erlang A Model, an arriving caller has a random time that they continue to wait and then abandon the queue - they do not wait indefinitely. For both models, no caller is blocked.*

The Erlang C Model is essentially an $M/M/n$ queuing system where there are n CSRs or agents independently providing statistically identical services. Calls typically arrive according to a Poisson process at a known mean rate of λ. Each CSR or agent takes a random (service) time to process a call; the service time follows an exponential distribution with parameter μ so that mean of the service times is μ.

Definition 3 (Offered load) *The offered load, denoted by R,*

$$R = \frac{\lambda}{\mu}. \tag{2.1}$$

Definition 4 (Offered Utilization) *The offered utilization (or simply* **utilization***, or* **traffic intensity***) of the queuing system or call center is denoted by ρ:*

$$\rho = \frac{\lambda}{n\mu} = \frac{R}{n} \tag{2.2}$$

Remark 2 *The offered load (of an agent or CSR) is a unit-less quantity which gives the number of Erlangs per agent while the offered utilization represents the proportion of available agent time spent handling calls if an assumption is made that all calls are processed. For queue stability, $\rho < 1$.*

The following results (related to Key performance indicators/measures of a queuing system) apply to an Erlang C Model:

1. The probability of an incoming call having to wait (i.e. the probability that all CSRs or agents are busy), denoted by $P(\text{Wait} > 0)$;

$$P(\text{Wait} > 0) = 1 - \left(\sum_{m=0}^{n-1} \frac{R^m}{m!}\right) \div \left(\sum_{m=0}^{n-1} \frac{R^m}{m!} + \left(\frac{R^n}{n!}\right)\left(\frac{1}{1-R/n}\right)\right) \tag{2.3}$$

2. The average speed to answer for a call center, denoted ASA:

$$\text{ASA} = E[\text{Wait}] = P(\text{Wait} > 0).\left(\frac{1}{n}\right).\left(\frac{1}{\mu}\right).\left(\frac{1}{1-\rho}\right) \tag{2.4}$$

3. The telephone service factor (TSF), sometimes referred to as the service level, is that proportion of processed calls for which the delay is below a specified level. If the specified level is T_o time units,

$$\begin{aligned} \text{TSF} &= P(\text{Wait} \leq T_o) = 1 - P(\text{Wait} > 0).P(\text{Wait} > T_o)|\text{Wait} > 0) \\ &= 1 - P(\text{Wait} > 0).e^{-n\mu(1-\rho)T_o} \end{aligned} \tag{2.5}$$

2.4.2 Erlang B Model

At its inception, the Erlang B Model was developed for the fixed network. According to Kendall's notation, it is an $M/M/n/n$ queuing model, i.e. the queuing system has n CSRs or agents and can accommodate a maximum of n customers - any arriving customer who arrives when all the CSRs or agents are busy is turned away or blocked. Qiao and Qiao (1998) give a good account of the attributes of the Erlang B model and present what they call *a robust and efficient algorithm for evaluating Erlang B formulae*. Tunnicliffe et al. (1998) have conducted a statistical analysis on cellular network calls data and come to the conclusion that the Erlang B model correctly modeled the data (at a significance level of 0.05) when the number of channels is in excess of 12 and the blocking experienced is greater than 1%.
The model assumes that

- the number of users is large;
- the customer arrival process is a Poisson process with known mean arrival rate λ;
- the customer service times are exponentially distributed;
- there is full availability (i.e. an arriving call can be processed by any available CSR)
- lost calls are cleared (i.e. they leave the system and do not attempt re-entry)

Remark 3 *Any call arriving at the call center when all the CSRs or agents are busy will be blocked. For this reason, the model is called a* **loss system**.

Definition 5 (Erlang Loss Function) *The mathematical function*

$$B(n,x) = \frac{\frac{x^n}{n!}}{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}} \tag{2.6}$$

is called the Erlang loss function.

Note that in the above definition n stands for the number of agents and x would represents the offered load.

The Erlang Loss Function is useful in the computation of the probability that an arriving call will be blocked for a Call Center operating according to the Erlang B model assumptions.
The probability of blocking is given by:

$$p_n = \frac{\frac{R^n}{n!}}{\sum_{i=0}^{n} R^i/i!} = B(n, R) \tag{2.7}$$

Remark 4 *It is remarked that the direct computation of p_n especially when R and n are large using Equation (2.7) (sometimes referred to as the* **service grade***) is difficult*[2] *and the recommended route is to use Equation (2.8)*

$$B(n, R) = \frac{RB(n-1, R)}{n + RB(n-1, R)} \text{ and } B(0, R) = 1 \tag{2.8}$$

Remark 5 *For a call center where the Erlang B model is applicable, the offered load R is given and the challenge is that of finding the minimum number of CSRs needed to achieve a desired industry service grade p_o, i.e. the minimum number n such that $p_n = B(n, R) < p_o$.*

2.4.3 Erlang A Model

A number of recent papers have advocated use of the Erlang A Model acclaiming it as a more accurate representation of call center operations. For instance, Gans et al. (2003) point out that call abandonment cannot be ignored, especially in the case of a high volume call center. They recommend the use of the Erlang A model as the *standard* to replace the popular Erlang C model.

The Erlang A model, also known as the $M/M/n + M$ queuing system, is the simplest model that incorporate the effect of caller impatience resulting in call abandonment. According to Knessl and van Leeuwaarden (2015), the Erlang A model is used in practice and is studied because it provides important approximations to general queue abandonment models.

Mandelbaum and Zeltyn (2007) and Knessl and van Leeuwaarden (2015) explore the Palm/Erlang A Queuing Model and its applications to a call center

[2]Qiao and Qiao (1998) mention that this stems from R^n growing very fast - e.g. if $R = 100$, then for $c > 154$, R^c approaches to ∞

where call abandonment cannot be ignored. According to Mandelbaum and Zeltyn (2007), the model has four parameters, which are:

- λ - the mean arrival rate of calls (per unit of time); call arrivals are assumed to be Poisson distributed with parameter λ so that the inter-arrival times of calls follow an exponential distribution with parameter $\frac{1}{\lambda}$;
- μ - the (mean) service rate (so that the average service time is $1/\mu$); the service times follow an exponential distribution with parameter $\frac{1}{\mu}$;
- m - number of parallel servers or CSRs;
- θ - the individual impatience parameter or abandonment rate (so that the waiting customers abandon the system[3] with an average *patience time* that follows an exponential distribution with parameter θ)[4].

Remark 6 *The Erlang A model assumes independence between inter-arrival times, service times and reneging times.*

Remark 7 *The queue length at time t is denoted by $N(t)$ and the queue length process $(N(t))_{t\geq 0}$ is a pure birth-death process. The birth and death rates when $N(t) = j$*

$$\lambda_j = \lambda \text{ and } \mu_j = \begin{cases} j\mu, & j \leq m \\ m\mu + (j-m)\theta, & j > m \end{cases} \tag{2.9}$$

Knessl and van Leeuwaarden (2015) give derivations of explicit expressions for the Laplace Transform of the time-dependent distribution and the first passage time. Some of the key theoretical results in Knessl and van Leeuwaarden (2015) include

- if $\theta = 1$ the Erlang A becomes the $M/M/\infty$ model, i.e. the model with an infinite number of CSRs;
- As $\theta \to 0^+$ the Erlang model approaches $M/M/m$;
- As $\theta \to \infty$, the $M/M/n/n$ model is approached.

[3] a customer who abandons the queue may be referred to as a reneging customer

[4] in the paper by Knessl and van Leeuwaarden (2015), they use η in place of θ

2.4.4 Erlang R Model

The discussion of the Erlang R model in this research report is simply for completeness's sake. An inspection of caller numbers revels that there is a low incidence of repeat calls and for this reason the Erlang R model is not appropriate for the data.

Remark 8 *The Erlang R model fills in a void left by the Erlang A, B and C models regarding, according to Yom-Tov and Mandelbaum (2014), "customers who return for service several times during their sojourn within the system".*

In the Erlang R model developed by Yom-Tov and Mandelbaum (2014), it is assumed that a caller who has received service will exit the call center system with probability $1 - p$ and they return for service after a random delay time with probability p. The arrival process is assumed to be a time-inhomogeneous Poisson process with rate function λ_t.

2.4.5 Hyper-Erlang Model

The Hyper-Erlang model is discussed by Fang (2001); the author asserts that it should be the natural choice for tele-traffic modeling in communication networks with integrated services - the call center environment fits this description very well.

The model assumes that the queuing system supports M types of services. Call requests of Type i are assumed to be a Poisson arrival stream with rate $\lambda_i,\ i = 1, \cdots, M$ while $s_i(t)$ is used to denote the service time of call requests of Type i.

2.4.6 Limitations and strengths of Erlang Models

The criticism of the Erlang C model is a result of the assumptions made:

- Assumption of known mean arrival rate of calls - the validity of this assumption arouses a lot of debate
- Assumption of no call abandonment - this has been widely criticized as being not in conformity with a typical call center environment

- Assumption of service times following an exponential distribution - a breakdown in this assumption leads to the correct model being $M/G/n$ as opposed to it being $M/M/n$ and the $M/G/n$ is said to be analytically intractable.

The effect of requesting for a specific service multiple times and even re-entering customers during their stay within the system is not taken into account by any of the Erlang models considered above. In emergency healthcare environments for instance, a customer may require a particular service or resource several times. To this end, the customer(s) re-enter the queue facility and request for the service repetitively (see Zhan and Ward (2013)).

2.5 Call arrival process an inhomogeneous Poisson process

Some research work has noted that the arrival process of calls at a call center varies over time and hence making the process an in-homogeneous Poisson process. A summary of the papers relating to this is given in this section.

Avramidis et al. (2004) have developed stochastic models that take into account a departure of call arrival processes at call centers from the classical assumptions. In particular they mention three unusual properties relating to the call arrival process at call centers that have been observed from recent empirical studies:

[P_1] the variance of the number of calls arriving per unit of time has tended to be way more than the mean of the number of calls and they refer to this as over-dispersion. Over-dispersion then casts doubt.

[P_2] the call arrival rate at a call center is time-dependent

[P_3] call arrivals in non-overlapping time intervals are not uncorrelated.

They go on to mention that the classical well-known non-homogeneous Poisson process (NHPP) is not consistent with the first properties, P_1 and P_2.

Consequent to the first property (P_1), Jongbloed and Koole (2001) have suggested the modeling of the arrival rate using a doubly stochastic Poisson model so that the arrival rate is a Gamma distributed random variable.
As a way of dealing with properties P_2 and P_3, Whitt (1999) has proposed the use of a doubly stochastic Poisson process with the arrival rate function over a day of the form $\Lambda(t) = W(f(t))$, where $W \sim Gamma(\gamma, 1)$. He goes on to suggest that the analysis of the call arrival process be done by means of partitioning the time horizon into appropriately small non-overlapping and *exhaustive* time segments as explained subsequently. Let $\mathcal{P}$ be a partition of the interval (t_S, t_E) so that

$$t_S = t_0 < t_1 < t_2 < \cdots t_{k-1} < t_k = t_E$$

with mesh, δ defined as $\delta = \max |t_k - t_{k-1}| \to 0$ as $k \to 0$. (For a given day, δ could be of duration 15-30 minute.) Further, let $\mathbf{X} = (X_1 m \cdots, X_k)^T$ be a random vector with elements $X_1, \cdots, X_k$ denoting the number of arriving calls in the respective k time segments and $Y = \sum_{i=1}^{k} X_i$ be the total number of calls in the time interval (t_S, t_E).

Remark 9 *The cornerstone of the work in the paper by Whitt (1999) is that the joint distribution of $X_1, \cdots, X_k$ is a negative multinomial distribution:*

$$P(X_1 = x_1, \cdots, X_k = x_k) = \frac{\Gamma\left(\gamma + \sum_{i=1}^{n} x_i\right)}{\Gamma(\gamma) \prod_{i=1}^{n} x_i!} \left[1 + \sum_{j=1}^{n} \lambda_j\right]^{-\gamma} \prod_{i=1}^{n} \left(\frac{\lambda_i}{1 + \sum_{j=1}^{n} \lambda_j}\right)^{x_i} \tag{2.10}$$

The paper by Liao et al. (2012) explores the problem of staffing a call center in a multi-period time frame with each period treated as a segment with a uniform call arrival rate. They begin by noting that, traditionally a call center has been assumed to have a known time independent and constant call arrival rate - mainly because this results in tractability. They go on to note that this does not apply to call centers where parameters such as call arrival rate has an element of uncertainty.

Remark 10 *Liao et al. (2012) model the call arrival process as a doubly non-stationary stochastic process, with random mean arrival rates.*

This paper has not been discussed in depth here because tests for stationarity in the call arrival process have not been supportive of the process being non-stationary.

2.6 Call arrival process a stationary process

One important aspect to be checked probably before anything else is that of whether the call arrival process is stationary or not. The augmented Dickey-Fuller test (ADF) is employed for this purpose.

A time series whose statistical properties such as mean, variance, auto-correlation, etc. are all constant over time is said to be stationary. The ADF statistic is normally a negative number, and the more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.
The unit root null and stationarity hypotheses to be tested in checking whether sample realized data is stationarity or not:

H_0: $\pi = 0$ versus

H_1: $\pi < 0$, respectively.

Here $\pi = \theta_1 + \theta_2 - 1 = -\theta(1)$ and $c = -(\theta_1 + \theta_2)$ are obtained from the AR(p) regressive model (with p = 2) for the process Z_t:

$$\begin{aligned} Z_t - Z_{t-1} &= (\theta_1 - 1)Z_{t-1} + \theta_2 Z_{t-2} + \xi_t \\ &= (\theta_1 + \theta_2 - 1)Z_{t-1} - \theta_2(Z_{t-1} - Z_{t-2}) + \xi_t \\ &= \pi Z_{t-1} + c(Z_{t-1} - Z_{t-2}) + \xi_t \end{aligned}$$

Setting $\Delta Z_t = Z_t - Z_{t-1}$ then we get

$$\Delta Z_t = \pi Z_{t-1} + c\Delta Z_{t-1} + \xi_t \tag{2.11}$$

where ξ_t stands for the stochastic part which is the economics innovations that can be accumulated to a random walk usually known as noise or independent identically distributed real valued random variable also called standard Brownian motion (Wiener Process).

The ADF test statistic for the test is $\hat{\tau} = \dfrac{\hat{\pi}}{\mathbf{se}(\hat{\pi})}$ where $\mathbf{se}(\cdot)$ is the standard error function.

Remark 11 *Equation 2.11 results in the general regressive AR(p) model for the ADF given by*

$$\Delta Z_t = \theta Z_{t-1} + c_1 \Delta Z_{t-1} + c_2 \Delta Z_{t-2} + \cdots + c_p \Delta Z_{t-p} + \xi_t. \tag{2.12}$$

The hypotheses correspond to

H_0*:* $\pi = 0$ *i.e., in order to make the process* Z_t *leading the data stationarity, the process needs to be differenced; versus*

H_1*:* $\pi < 0$*, i.e., the process* Z_t *leading the data is stationarity, and there is no need for the data to be differenced.*

Chapter 3

Methodology

3.1 Introduction

This chapter unpacks in detail different processes and techniques used in the analysis of a university call center data. The first stage consists of exploring the data used in the analysis. This exploration specifically focuses on the visualization of call center operations data using descriptive statistics which give the reader an overview of the methods used in the analysis of the call centre data and hopefully give a better understanding of the queuing system.

3.2 The Call Center Data

The data obtained from the university inbound call center spans more than one year (438 days); from the 3rd January 2017 to 3rd October 2018. The call center operates daily from Monday to Friday; from 8:00 to 16:30, and on Saturdays from 8:00 to 12:00. There were 109047 calls that entered the system during this time frame. Customers call in and the router system routes the calls to the relevant agent or queue. Once directed to an agent, the agent assists the customer if they are able to; otherwise the call is re-routed to someone else who can assist. The data are then stored on the database. Among other things, the call center system records the following:

- the arrival time and date for each call, t_i (one can then be able to calculate the inter-arrival times for two successive calls whose arrival times are t_{i-1} and t_i which would be $t_i - t_{i-1}$),

- the time at which a caller starts to receive service, τ_i so that the waiting time for caller i is $\tau_i - t_i$
- the time at which Caller i's service is completed, t_i' for purposes of calculating the caller service time (t_i' - τ_i)
- the number of arriving calls which are blocked in a given period of time (useful when the Erlang B model applies).
- etc.

The analysis of customers arrival process involves dividing time into non-overlapping 15-minute segments and analysing the numbers of customers who arrive in each segment. This can also enable analysis of abandonment (when Model A is used)

3.3 Limitations / Challenges in obtaining quality call center data

The major problems usually faced in obtaining good quality call center data include:

1. Confidentiality and Protection of Personal Information (POPI) laws continue to present a major limitation on accessing the data. Management of a call center exercise extreme caution in releasing data to outsiders in order to protect personal information.[1] This restricts the amount and kind of data we can extract and use and also restricts the kind of analysis that can be performed.
2. Missingness of data - in the unfortunate event of missing data in the dataset, presumably because of problems such as Information Technology (IT) system shut down due to maintenance, recording errors, or any other associated glitch, an appropriate imputation method could be considered for use to deal with the missingness. In dealing with data cleaning and missingness of data as well as manipulation and analysis,

[1] A number of processes such as getting ethics clearance, assuring the management that utmost care will be taken in the use of the data, etc. had to be done before acquiring the data.

R (statistical software for computing) and/or SPSS are usually sufficient.

3.4 Description of variables

Table 3.1 gives the detailed description of the data fields.

Table 3.1: Preliminary Data background

Period	01/01/2017 – 03/10/2018 08:00 – 16:30.
Agents	1, 2, 3, 4, 5, 6.*
Starter Date/Time	Arrival time/date of a call in the system, i.e. time/date the call arrived in the Starter.
Router Date/Time	Time the call arrived at the router queue.
Caller	The originator of a call identified by their telephone number ***).
Router	The name of the router as entered during the creation of the router.
Response Time	Time the caller spends on the router including agent alerting time.
Call Duration	Time between answering of a routed call and ending the routed call.
Transfer Destination	Phone number of the agent a call is transferred to.
Time segment	Time is divided into segments each of 15 minutes duration.

2 3 4

[2]* Agents are employees of the university call center who are identifiable by their employee numbers; here they are disguised as $1, \cdots, 6$

[3]** In the workplace, these are employees at the university call centre and they are identifiable through their employee numbers.

[4]*** Not made use of to ensure conformance to confidentiality requirements

Offered Calls	Number of routed calls offered to this router.
Answered	Number of answered routed calls to this router.
Transferred	Number of routed calls that were answered by an agent and then transferred to another party.
Callback	The number of times callers stopped waiting in a queue via a callback request.
Callback retain position	Number of times the caller stopped waiting in the queue and requested a callback with retaining queue position.
Abandoned	Number of routed calls that were disconnected by a caller before they were answered.
No Agent	Number of routed calls rejected or re-routed because of a no agent situation (all agents "not ready" or "logged off").
Queue Full	Number of routed calls that were rejected or re-routed because of queue full situation.
Queue Abort	Number of routed calls that were aborted via an option menu by a caller before they were answered.
Queue Timeout	Number of routed calls rejected or re-routed because of queue timeout.
Ready Agents	Average number of agents switched "Ready" at the moment each call was queued to the router.
Not Ready Agents	Average number of agents switched "not ready" at the moment each call was queued to the router.

3.5 Queuing System Primitives

To successfully run a call center, managers have to adhere to a well fine-tuned balance between the efficiency of the *CSR* or agent and the quality of the service provided and to fulfill this purpose, the use of mathematical models to sustain queuing theoretic models is inevitable. The inputs to these models, called *queuing system primitives* are the number of *CSR*s, the calls arrival rate, customer service time, the time a customer is willing to wait before abandoning the queue.

These stochastic queuing system primitives are examined in the first phase of the analysis. Section 3.6 discusses the methodology of analyzing the calls

arrival process, while Section 3.8.1 explores the service times and Section 3.9 deals with the methodology employed in analyzing call waiting times for service and the issue of call abandonment.

3.6 Calls Arrival Process

Customers of the call center are considered to be in queue and are served according to the first-come first-served (FCFS) queue discipline, and they are distinguished by their caller identity (which is their telephone numbers) and their arrival time. In some systems, while customers are waiting, from time to time they receive information regarding their progress in the queue.

Usually the arrival of calls constitute a Poisson process - which can be homogeneous or inhomogeneous.

Definition 6 (Homogeneous Poisson process) *A counting process* $\{N_t, t \geq 0\}$ *is said to be a Homogeneous Poisson process having rate or* **intensity** $\lambda > 0$*, if*

1. $N(0) = 0$.
2. *The process has independent increments.*
3. *The number of events in any interval of length t is Poisson distributed with mean* λt*. Therefore* $E[N(t)] = \lambda t$.

Adams et al. (2009) says that an inhomogeneous Poisson process is a point process with a varying intensity across its domain (which may be time or space).

Remark 12 *In layman terms, the main distinction between a homogeneous process and an in-homogeneous process is that, for the former, the intensity is not time-dependent while for the latter the intensity is time-dependent.*

The characterization of a non-homogeneous or in-homogeneous Poisson process or non-stationary Poisson process given in Definition 7 is taken verbatim from the book by Ross (2014) (pages 284-6).

Definition 7 (Non-homogeneous Poisson process) *A counting process* $\{N(t), t \geq 0\}$ *is said to be a non-homogeneous Poisson process with* **intensity** $\lambda(t),\ t \geq 0$*, if*

1. $N(0) = 0$.

2. $\{N(t), t \geq 0\}$ *has independent increments.*

3. $P[N(t+h) - N_t \geq 2] = o(h)$

4. $P[N(t+h) - N_t = 1] = \lambda(t)h + o(h)$

Remark 13 *Brown et al. (2005) assert that it is common practice to model a call centre arrival of calls as a homogeneous Poisson process with a rate that remains constant in each time segment; the rates may however be different from one segment to another. With such an approach, the arrival process is described by a rate function can be considered to be well approximated by a piecewise continuous function.*

Brown et al. (2005) constructed a test that can be used to check if a stochastic process is an inhomogeneous Poisson process with rates that are piecewise constant. The procedure involves

- breaking up a day into short non-overlapping intervals of time or segments (say 15 minutes)

- zero in on a block of time for different days - in the i^{th} block. If T_{ij} denotes the j^{th} ordered arrival time in the block, and

$$R_{ik} = (J + 1 - k)\left(-\log\left(\frac{L - T_{ik}}{L - T_{i,k-1}}\right)\right);\ k = 1, \cdots, J \qquad (3.1)$$

  where J is the total number of calls received in the block and L is the block length (15 minutes), then R_{ik} will be independent standard exponential variables as described in their paper.

Remark 14 *To test conformity to a distribution, customary statistical tests such as the chi-square goodness-of-fit test, one-sample Kolmogorov-Smirnov and Anderson-Darling (discussed in Section 3.7) can be carried out. Other complementary tools include visual techniques such as the quantile-quantile (QQ-plot) can be used to ascertain the distribution.*

3.7 Goodness of fit tests

3.7.1 Chi-square goodness of fit test

The Chi-square goodness of fit test has the advantage that it is usable for both discrete and continuous probability distribution. In conducting the Chi-square goodness of fit test, the sample data is divided into intervals or classes. The frequency of each class or number of points that fall into each class (f_i) is then noted and compared with the expected numbers of points in each class (e_i) The expected number of points (expected frequency) is computed using the hypothesized distribution.

The hypotheses to be tested are

$H_0 : X \sim \mathbb{P}(\cdot)$ vs

$H_1 : X$ does not follow the distribution $\mathbb{P}(\cdot)$

The test statistic, χ^2, is calculated by:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{3.2}$$

where the decision to reject H_0 is dependent on whether or not χ^2 value is less than or greater than some critical value obtained from statistical tables. The decision is equivalently made if the p-value from the test is less than the set level of significance.

Remark 15 *The Chi-square goodness of fit test has the advantage that one does not need to completely specify the values of the parameters of the hypothesized distribution in order to conduct the test.*

A more detailed account of the test is given by Pearson (1916).

3.7.2 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov Goodness of fit test (K-S test) is a non-parametric test used for determining whether given sample data follow or do not follow

a hypothesized probability distribution.

The null and alternative hypotheses to be tested in checking whether sample realized data follow a distribution $F(\cdot)$ are:

H_0: sample data follow $F(\cdot)$ versus

H_1: sample data do not follow $F_0(\cdot)$, respectively.

The computation of the test statistic involves first computing sample cumulative distribution function or empirical distribution function values of the sample data $F_{\text{data}}(x_i)$; the test statistic

$$D = \sup |F_{\text{data}}(x_i) - F_0(x_i)| \tag{3.3}$$

is compared with Table K-S values to decide whether H_0 is rejected or not.

Remark 16 *one major importance of the K-S test is that it works fine regardless of sample size. Its major limitation is that it cannot, in general be used for discrete distributions.*

Lilliefors (1967) gives a detailed account of the Kolmogorov-Smirnov statistical test of hypothesis.

3.7.3 Anderson-Darling goodness of fit test

The Anderson-Darling (AD) test, just like the K-S test, is a goodness of fit test used for determining whether a given sample data follows or does not follow a hypothesized probability distribution.

The null and alternative hypotheses to be tested in checking whether sample realized data follow a distribution $F(\cdot)$ are:

H_0: Sample data follow $F_0(\cdot)$ versus

H_1: Sample data do not follow $F_0(\cdot)$, respectively.

The test statistic for the test is

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F_0(x_i) + \ln(1 - F_0(x_{n-i+1}))] \qquad (3.4)$$

where the decision to reject H_0 is dependent on whether or not AD is less than or greater than some critical value obtained from statistical tables. The decision is equivalently made if the p-value from the test is less than the set level of significance. For a more detailed account, the reader is referred to Anderson (2011) and Anderson and Darling (1954).

3.7.4 Goodness of fit test based on the sample Gini index

For testing that sample data follow an exponential distribution, Gail and Gastwirth (1978) have developed an arguably very good goodness-of-fit test based on the sample Gini index. They say the sample Gini index is shown to be a powerful scale-free test of exponentiality against a variety of alternatives. The advantages of the test are

- Its good power compared to competing tests,
- Ease of computation of the test statistic,
- Robustness to measurement error, and
- Availability of exact critical values.

3.8 Operating characteristics of the call center

Computation of the operating characteristics of the call center (in line with Section 1.3) such as

1. mean waiting time, W,
2. mean number of customers in the queue or mean queue length = L_Q,
3. mean number of customers in the system = L,

4. probability of blocking

5. etc.

using various procedures proposed in a number of papers reviewed in Chapter 2, Koole and Mandelbaum et al. (2000), Robbins et al. (2006), Cruz et al. (2005), Gans et al. (2003) is done. Recommendations for attaining the call center service industry are made.

3.8.1 Service Times

The quality of the service given by a call centre is measured through different queuing metrics. One of the queuing metrics is based on service times for callers.

According to literature, most call center queuing practitioners assume that, by the default, the service times are exponentially distributed. The models used for call center data that assume an exponential distribution for service times usually give tractable result.

3.9 Call Waiting Times for Services

The calls arrival process and caller service times undoubtedly have an impact on how long callers wait for service as well as the incidence of abandonment.

Abandonment and call waiting are deeply entangled phenomena. A distinction between the two need to be made. There are two cases:

1. a caller who waits until they are routed to an available agent to receive service and they actually get the desired service

2. a caller who waits up to sometime and then gives up (in frustration) and leaves the queuing system before receiving service

For Case 1, the actual waiting time (virtual waiting time) to receive service is known. However, we cannot measure "patience". For Case 2, while we can measure "patience", we however, cannot measure the virtual waiting time.

Remark 17 *In both cases, there is an element of right censoring although the right censoring is for different variables. If the time a customer is willing to wait (which is a measure of patience) and the "virtual" waiting time (actual waiting time) of a customer are denoted by W' and V, respectively, then waiting time $W = \min\{W', V\}$.*

Ideally, a manager of a call centre will wish that customers wait until they receive service and in such a case the waiting times are usually adequately modeled by an exponential distribution. However, in reality customers are not willing to wait indefinitely and abandonment does inevitably occur.

3.9.1 Waiting time and patience survival curves

Survival analysis is a subject comprising a set of tools that are used to model the time it takes for an *event of interest* to occur. Survival time or "time to event" is the time to the occurrence of the event of interest. In this case there are two distinct events of interest: 1) time when service commences, .i.e. when a caller first talks to an agent and 2) time when a caller abandons the queue.
Thus, waiting time in the data set is a primitive which consists of two components: abandonment time and the time to service. These two components may apparently be both censored.

Indeed the analysis of this kind of data perfectly fits the realm of the subject area called Survival Analysis. Brown et al. (2005) decomposed the waiting time into two components, namely the "virtual" waiting time and the time "willing" to wait. Brown et al. (2005) correctly points out that what can be obtained for each call from call center data is a value of W and an indicator variable $I_{W'<V}$ which will take a value of 1 if $W' < V$ and 0 otherwise.

Remark 18 *According to Brown et al. (2005), all calls reaching the agent result in censored observations for W' while all calls that do not reach the agent are censored observations for V. In this work, the same approach of assuming W' and V to be independent random variables made by Brown et al. (2005) is taken and the standard Kaplan -Mier product limit estimator is made use of in estimating the cumulative distributions of W' and V.*

To better understand different Survival Analysis techniques for analysing waiting times, a discussion of a counting process paradigm is essential.

3.9.1.1 Patience and Impatience and Hazard rates

Palm (1953) is credited as the pioneer of modeling abandonment using a *hazard function*.

Definition 8 *If the time to failure is modeled by probability density function $f(t)$ and the survival function is denoted by $R(t)$, The hazard rate at time $t > 0$, $h(t)$ is defined as*

$$h(t) = \frac{f(t)}{R(t)} \tag{3.5}$$

Palm (1953) postulated that the hazard rate of the time willing to wait is proportional to the customer's irritation due to waiting. In the case of customers who get to receive service, the hazard function predicts the amount of time until the event of interest occurs and the hazard rate is the expected number of calls that would be answered at the time.
Brown et al. (2005) use a nonparametric procedure given in Equation 3.6 to estimate the hazard rate in a time interval of length δ

$$\text{Estimate of } h(t) = \frac{\text{the number of calls answered during time } (t, t+\delta))}{[\text{number at risk at time t}] \times \delta}. \tag{3.6}$$

Hazard rates give us information regarding the time dependent behaviour of customers, for instance, when the hazard rate is high, there will be a high tendency to abandonment, whereas constant hazard rates are memoryless, i.e. the propensity to abandonment remains the same regardless of the past.

3.9.2 Customer Index Patience

The customer patience index is an important element in call centre operations. Different definitions of the customer index patience can be found in queuing literature. The one given by Brown et al. (2005) relates to the ratio of the mean time a customer may be willing to wait and the mean time the customer actually needs to wait. If the two random variables are independent and exponentially distributed, we have

$$\text{Patience Index} = \frac{P(V > R)}{P(V < R)}. \tag{3.7}$$

and an estimator of the Patience index is the Empirical Index given by

$$\text{Empirical Index } = \frac{\text{number of served}}{\text{number of abandoned}}, \tag{3.8}$$

Remark 19 *The numbers of callers served and callers abandoned can be easily calculated using the call centre data.*

3.10 Predicting the Number of Staff

As discussed, the common primitives of a queueing system are arrival rates, service time, service waiting time and abandonment. Estimates of the latter primitives are used in conjunction with a queuing model such as the Erlang-A to predict the ideal number of staff needed to achieve a high level of customer satisfaction at the call centre.

The "Erlang" package embedded in the R software contains functions ideal for call centre staffing computations. The calculations use the following parameters:

- Number of calls per unit of time: this is the number of incoming calls per unit of time (e.g. per hour, per day, etc)
- Unit time (30 minutes, hour, etc.)
- Average call duration (or Average Handling Time (AHT)): this is the amount of time, on average, that an agent will take to handle a call. This may include wrap-up time (which is time spent on things like paperwork) before an agent gets to take the next incoming call.
- Service level: if the service level desired is such that 90% of calls are handled in 15 seconds, the inputs become 90 and 15
- Average Patience time (or Average Time to call Abandon (ATA)): call abandons are calculated using the Erlang A formula - it assumes an average patience time.
- Maximum Occupancy: the maximum number of agents needed for optimal service,
- Shrinkage: this factor takes into account holidays, sickness, etc.

Chapter 4

Queuing System Analysis

4.1 Introduction

This chapter presents the results gotten from analysis of the call center data using the techniques discussed in Chapter 3. The R statistical software (using different packages such as *ggplot2*, MASS, survival, *tseries* for example (to name a few))is consistently used in conjunction with SPSS and Excel in the analysis.

The analysis proceeds in the following order:

1. a comprehensive interrogation of the calls arrival process to check whether it fits into the realm of an Inhomogeneous Poisson process, etc.,
2. a statistical analysis of the calls service times,
3. an analysis of caller waiting times for Service, and ultimately
4. predicting the number of staff required in the call center.

4.2 The data

Table A1 in Appendix A is a screenshot of the data used in the analysis (after the data cleaning process). The data cleaning process involved the following:

1. removing Saturdays data. The reason for this is that the volume of call traffic on Saturdays was found to be very low given that the call center opens for only four hours on Saturdays.

2. all weeks with public holidays and in which the number of days when the call center was open was less than 5 were also removed.

3. data calls arriving in the time period 16h00 - 16h30 were also removed.

4. all calls with less than 10 seconds service times were removed.

4.3 Calls arrival process

Figure 4.1 is a time series plot of the numbers of calls recorded hourly during each of the *normal* week days.

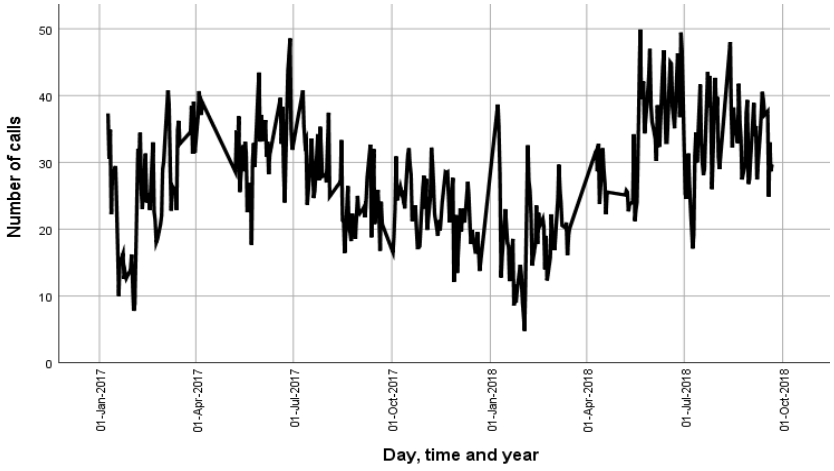


Figure 4.1: Distribution of Call by Hours

4.3.1 Stationarity of calls arrival process

In this section the Augmented Dickey-Fuller is used to test the data for stationarity. The following is the null hypothesis and its alternative.

H_0 variable contains a unit root / non-stationary
H_a Process is stationary

The t-series package is used to conduct the Augmented Dickey-Fuller test to check for stationarity in the hourly calls data (see Appendix B1).

The results (Dickey-Fuller = -7.9274, Lag order = 15, p-value = 0.01) reject the hypothesis that the hourly number of calls received are a non stationary process.

4.3.2 Distribution of calls over time

Figure 4.2 gives the comparative boxplots of the numbers of arriving calls for the different non-overllaping 1- hour periods[1] of the day for all the data.

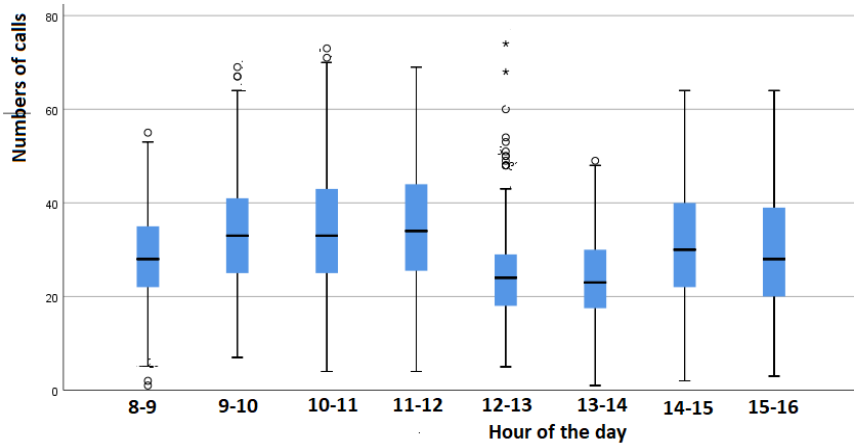


Figure 4.2: Comparative box-plots of numbers of calls per hour

[1] 8 represents the 08h00 to 09h00 time interval, 9 represents the 09h00-10h00, and so on

The bar chart in Figure 4.3 was produced using the *R-ggplot2* package.

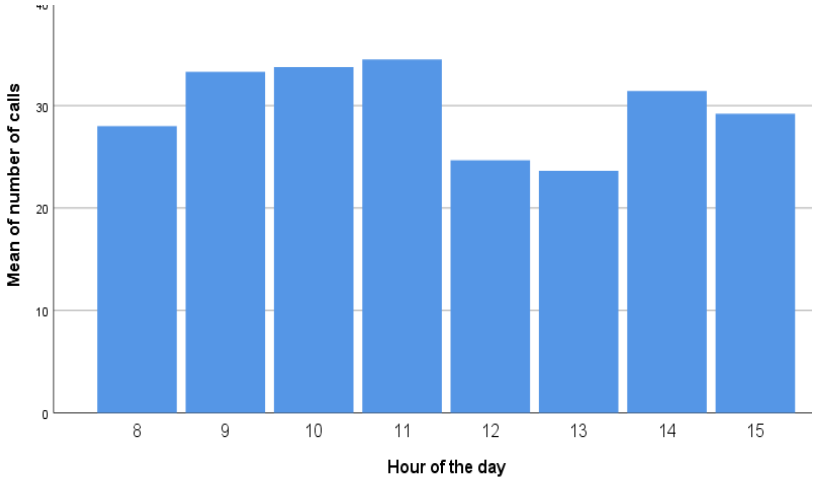


Figure 4.3: Comparative box-plots of numbers of calls per hour

Each bar in Figure 4.3 represents the mean of the number of calls for an hour long segment, with the first bar representing the first working hour, which starts at 08h00 and ends at 09h00 and so on till the last bar for the time segment 15h00 to 16h00. The time period 09h00 to 12h00 generally record higher volumes of calls than the rest of the times.

Table 4.1 (generated using R code) gives the numbers of calls recorded in each of the 1-hour segments.

Table 4.1: Numbers of calls per hour

8H	9H	10H	11H	12H	13H	14H	15H	Total
10047	11985	12119	12385	8830	8387	11248	10454	85455

For the entire data set, the maximum number of calls of 12385 per hour is recorded in the 11h00 to 12h00 time segment while the minimum of 8387 calls is recorded in the 13h00-14h00 time segment.

The distribution of arrivals of calls per day during the week was considered and is given in Figure 4.4 (after using the R code that can be found in Appendix B3.

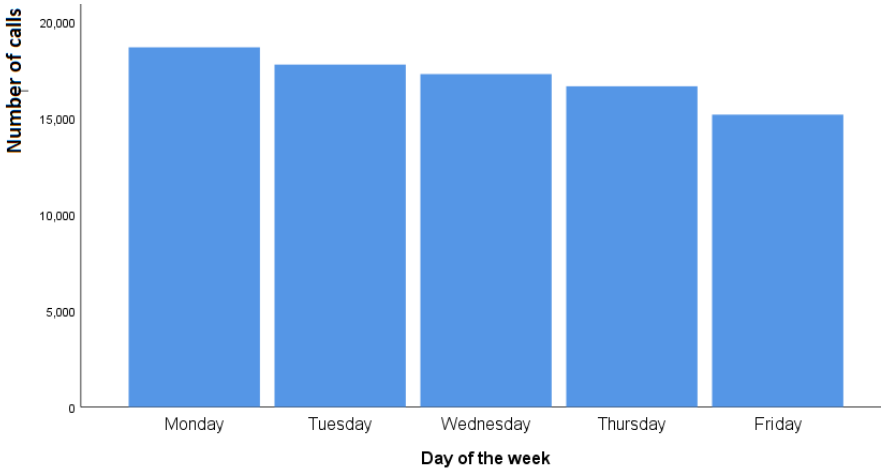


Figure 4.4: Distribution of Calls by Days of the Week

It is clear that Monday recorded the highest number of calls and the call numbers appear to slowly get smaller and smaller as the week progresses.

An inspection of the numbers of calls recorded per hour on a a randomly selected date (1st March 2018) was carried out. R code used yielded the results in Table 4.2 below and the actual code is put in Appendix B4.

Table 4.2: Distribution of Calls on 1st Day in March 2017

Time period	8H	9H	10H	11H	12H	13H	14H	15H
Number of calls	18	48	29	30	25	29	15	13

Interestingly, the distribution of the numbers of calls on the day shows a different pattern to that of the full data set; the maximum number of calls was recorded in the 1100 to 1200 hours time segment while the minimum was recorded in the 1500 to 1600 hours time segment for the full data set.

4.3.3 Call arrivals an Inhomogeneous Poisson process

From the discussion in the paper by Brown et al. (2005), it is imperative that one checks whether the numbers of calls that arrive in different time segments follow an inhomogeneous Poisson process with an arrival rate function which is piece-wise constant.

Initially, call arrival times recorded for each of the 8 time segments on the 9th of January, 2017 were analysed separately. For a segment, computation of the statistics r using the formula

$$r_j = (J + 1 - j)\left(-\log\left(\frac{3600 - t_j}{3600 - t_{j-1}}\right)\right) \tag{4.1}$$

where t_j=is the time in seconds it takes the j^{th} call to arrive after the commencement of the segment and J_i is the total number of calls recorded in the segment.

Histograms of the r_j values of each of the segments were constructed and they showed that the r values are typically exponentially distributed.

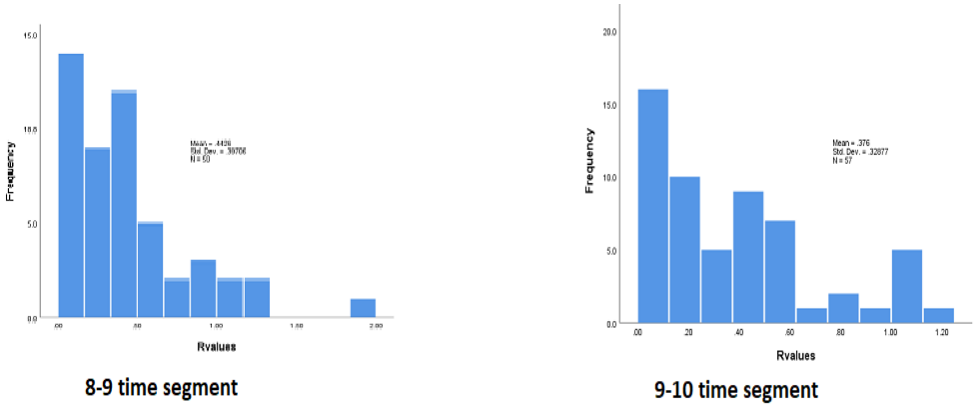


Figure 4.5: Histograms of r values from 8-9 and 9-10 time segments

The same process was repeated for many randomly selected time segments and all the rigorous test of hypotheses results suggested a lack of evidence

to reject the hypothesis that the r values in a segment were exponentially distributed. In each of the random samples, the p-value was greater than 0.05.

4.4 Service Time

This section is mainly focused with the analysis of of caller service times. A boxplot of the call service times show that the times are heavily skewed to the right and there is a high number of outlier service times.

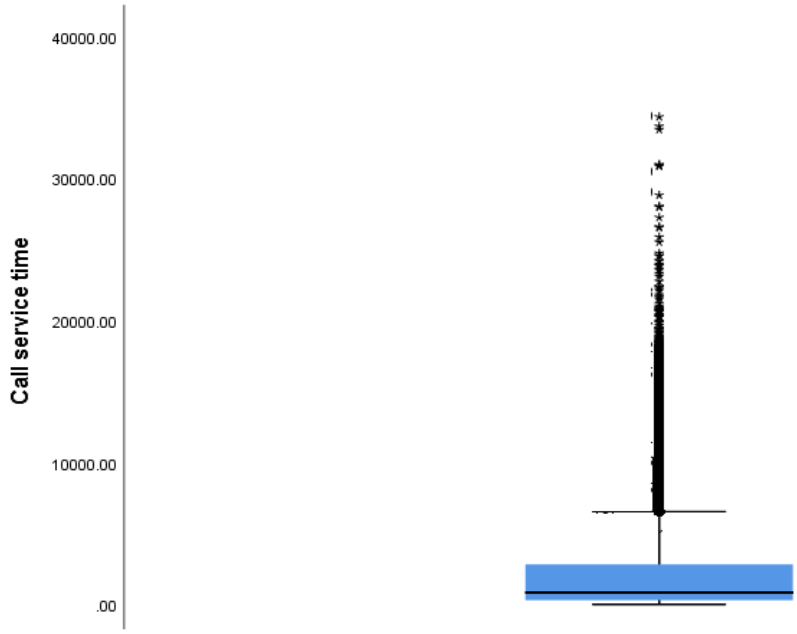


Figure 4.6: Boxplot of all call service times

The number of calls lasting more than 1 hour each was found to be 16799 (or 19.7%); most of which were found to be outliers and were therefore excluded in the analysis. Figure 4.7 is the histogram of the service times of all the calls lasting up to 1 hour; it reveals that service times do not follow an exponential distribution.

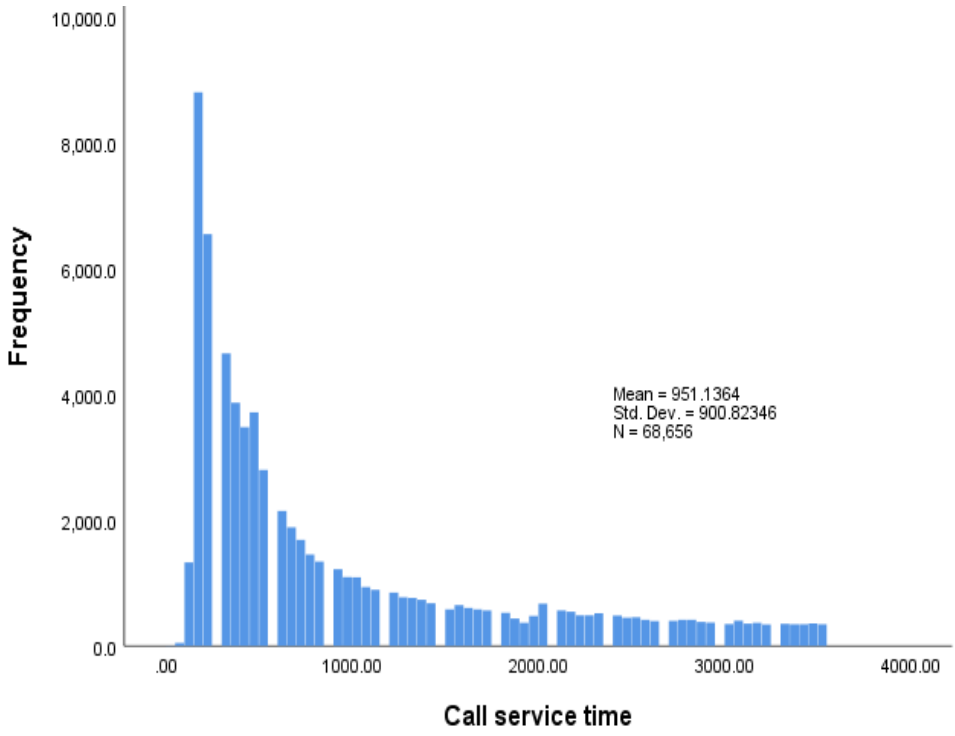
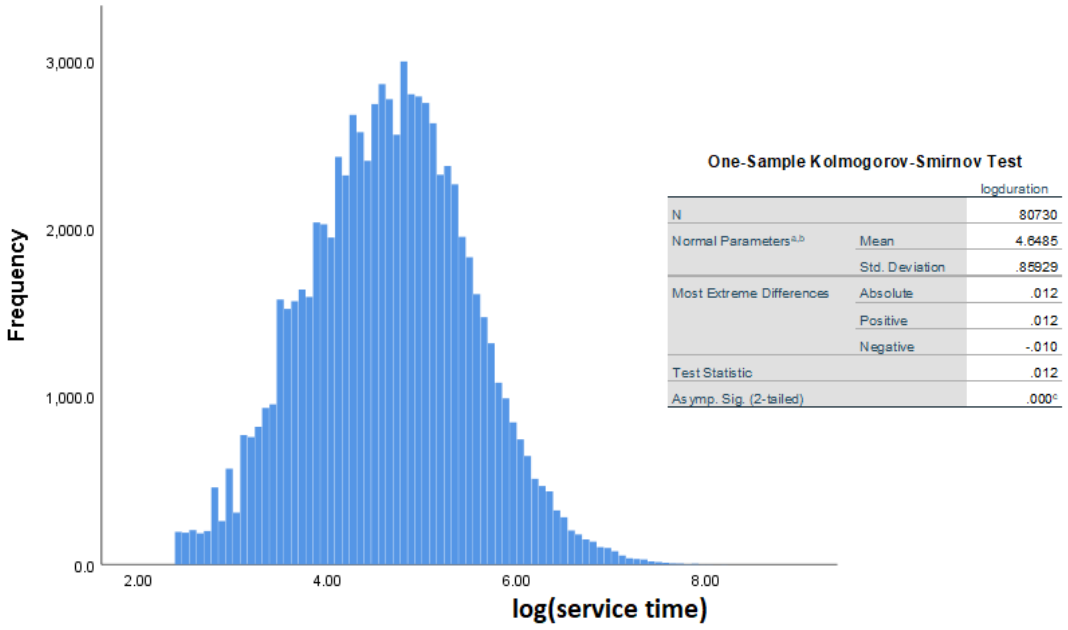


Figure 4.7: Histogram of call service times which are less than 1 hour

A histogram of the the log transformed service times (see Figure 4.8) seems to be supportive of the assertion that service times follow a log-normal distribution. However, the p-value in the Kolmogorov-Smirnov Test leads us to conclude that the service times do not follow a log normal distribution.



One-Sample Kolmogorov-Smirnov Test

		logduration
N		80730
Normal Parameters[a,b]	Mean	4.6485
	Std. Deviation	.85929
Most Extreme Differences	Absolute	.012
	Positive	.012
	Negative	-.010
Test Statistic		.012
Asymp. Sig. (2-tailed)		.000[c]

Figure 4.8: Distribution of Log Service Times

Figure 4.8 was produced using the below R code in Appendix B5.

The associated Q-Q plot (in checking the assertion of log-normal distribution of service times) in Figure 4.9 on the other hand contradicts the assertion.

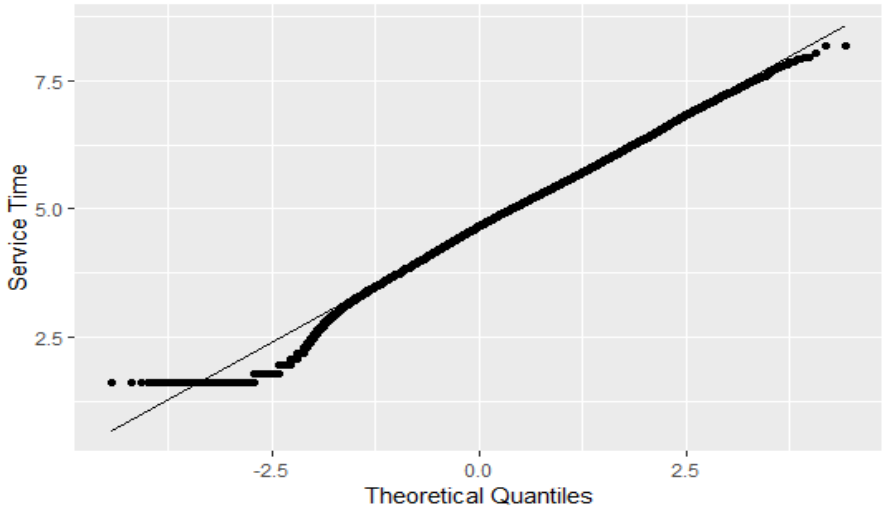


Figure 4.9: Q-Q Plot of the Log (Service Time)

The tests for normality firmly reject the null hypothesis of exact normality (for K-S statistic - p-value $< 2.2e{-}16$, and for Shapiro - p-value $= 3.59e{-}11$).

4.5 Waiting Times for Service

In the two preceding sections, an analysis of two primitives namely call arrivals and service times was done. It was established that arrivals are a time inhomogeneous Poisson process and service times are possibly log-normally distributed.

In this section, an investigation of customer patience and abandonment behaviors as well as the related waiting times is done.

These two notions, abandonment and waiting times are deeply entangled. However, there is a distinction between the time that a call needs to wait before it reaches an available agent and the time for the customer to wait until he gives up and leaves the system. The first case is referred to as waiting time and the second as patience. Both metrics are critical in queueing, but neither is truly observable and hence must be estimated.

In an ideal queuing system where all the customers can wait indefinitely and no one abandons the system, the waiting time should be exponentially distributed. Although the system under our investigation is not ideal, i.e. customers are not willing to wait indefinitely and surely there would be some abandonment, we find that the distribution of time customers spend in the queue aligns perfectly with theoretical prediction.

4.5.1 Waiting time and patience survival curves

The R code in Appendix B6 is used to produce the results that relate to the plots of customer caller patience times in Figure 4.10.

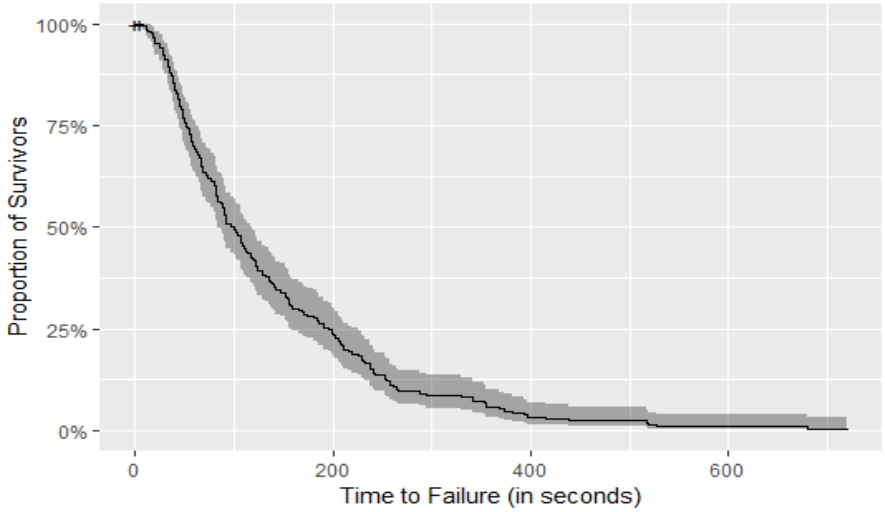


Figure 4.10: Survival Function of Patience Waiting Time

Figure 4.10 depicts the survival function giving the proportions of calls answered in the queue at a particular time t. The time to failure, here, translates to the time until a call will reach an agent or will leave the system for another reason. An inspection of the variation of the proportion of survivors with time appear to suggest that it is an exponential decay curve.

The second component obtained through the Kaplan-Meier estimates is the

virtual waiting time V, or abandonment, which is translated via the survival function in proportion and plotted against the time, as shown below.

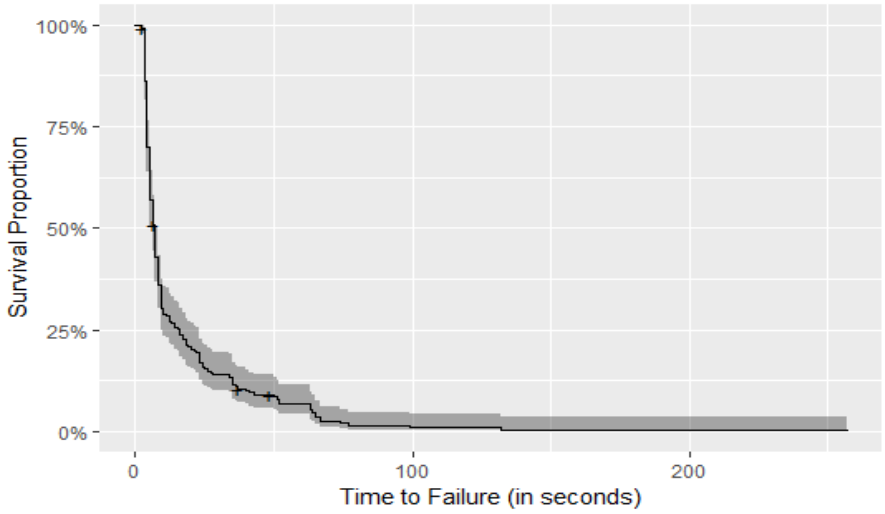


Figure 4.11: Survival Function of Virtual Waiting Time

Figure 4.11, above shows the proportion of calls that left the system after waiting for some time. The R code used is in Appendix B7.

Counting Process and the Hazard rates

The results in Table 4.3 is an output of survival analysis done for data on a randomly selected day using the R software with "survival" and the "survminer" packages. In the table, point estimates of probabilities of survival beyond time t are given as well as the "upper" and "lower" confidence limits of the 95% confidence interval.

Table 4.3: Survival outputs

time	Number at risk	Number of events	Number of censored	survival probability	Confidence limit	
					Lower	Upper
0	210	0	3	1	1	1
4	207	0	1	1	1	1
8	206	1	0	0.995146	1	0.985699
11	205	2	0	0.985437	1	0.969213
13	203	1	0	0.980583	0.999608	0.961919
16	202	1	0	0.975728	0.996971	0.954938
17	201	1	0	0.970874	0.994111	0.94818
18	200	1	0	0.966019	0.99108	0.941592
19	199	1	0	0.961165	0.987914	0.935141
20	198	2	0	0.951456	0.981261	0.922557
25	196	2	0	0.941748	0.974281	0.9103
27	194	1	0	0.936893	0.970693	0.90427
28	193	3	0	0.92233	0.959614	0.886495
29	190	1	0	0.917476	0.955831	0.880659
30	189	1	0	0.912621	0.95201	0.874862
32	188	1	0	0.907767	0.948153	0.869101
33	187	3	0	0.893204	0.936392	0.852008
35	184	3	0	0.878641	0.924384	0.835161
36	181	1	0	0.873786	0.920333	0.829594
37	180	2	0	0.864078	0.912167	0.818523
38	178	2	0	0.854369	0.903921	0.807533
39	176	1	0	0.849515	0.89977	0.802066
40	175	3	0	0.834951	0.887215	0.785766
41	172	1	0	0.830097	0.882998	0.780365
42	171	2	0	0.820388	0.874519	0.769609
43	169	1	0	0.815534	0.870257	0.764252
44	168	1	0	0.81068	0.865982	0.758909
45	167	3	0	0.796117	0.853079	0.742958

Table 4.3 (continued): Survival outputs

time	Number at risk	Number of events	Number of censored	survival probability	Confidence limit	
					Lower	Upper
46	164	1	0	0.791262	0.848752	0.737666
47	163	2	0	0.781553	0.840064	0.727118
48	161	3	0	0.76699	0.826948	0.71138
50	158	2	0	0.757282	0.81815	0.700942
52	156	2	0	0.747573	0.809311	0.690544
53	154	1	0	0.742718	0.804877	0.68536
54	153	2	0	0.73301	0.79598	0.675021
55	151	1	0	0.728155	0.791518	0.669865
56	150	2	0	0.718447	0.782566	0.659581
57	148	2	0	0.708738	0.773579	0.649332
58	146	2	0	0.699029	0.764557	0.639117
60	144	1	0	0.694175	0.760034	0.634022
61	143	1	0	0.68932	0.755503	0.628936

The following are applicable to the data in the table:

- time - denotes time point when the curve has a step,
- Number at risk - this is the number of subjects who are at risk of death at time t,
- Number of events - this represents the number of events occurring at the time t,
- Number censored - This is for the number of observations that are censored,
- Survival probability - this is an estimate of the probability of survival beyond time t,
- upper and lower - these are the upper and lower confidence limits for the probability of survival

Figure 4.12 is an extract of the cumulative hazard rate function obtained from the survival function. The hazard function is a monotonic increasing function. The confidence interval of the hazard function apparently gets wider over time.

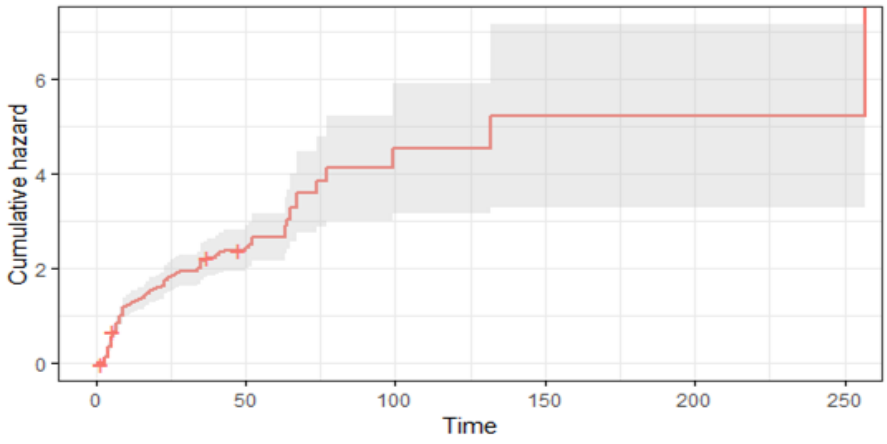


Figure 4.12: Hazard Cumulative Function

4.6 Predicting the required number of Staff Loads in the call center

After calculating estimates of the common primitives of a queuing system, namely, arrival rates, service times, and the service waiting times plus abandonment, the next step is to use the results to get estimates of the characteristics in a proposed model.

Remark 20 *In this research report, the use of the Erlang A and the Erlang C models is explored for purposes of determining the better of the two. It should be noted that for Erlang A, caller abandonment is a additional primitive of the queuing system that is necessary, while in the case of the Erlang C there is no provision for call abandonment.*

In both the Erlang A and Erlang C models, the call queueing system primitives can be used to predict the ideal number of staff needed to achieve a

high level of customer satisfaction for the call centre.

The "Erlang" package from the R software contains functions ideal to call centre staffing. It is made use of to calculate the required number of operators as explained in Section 3.10.

4.6.1 Fitting the Erlang A model

An analysis of the call center data for Tuesday (Tuesday is the day with a peak in terms of numbers of calls) was done. The analysis was also repeated for a randomly selected day.

Results for a randomly chosen day

Table 4.5 gives the results that were obtained. In the table, SL stands for Service Level, ABS stands for Agents Before Shrinkage, AI stands for the percentage of calls Answered Immediately, AR stands for call Abandonment Rate and N the number of agents.
In Table 4.5, for instance, in row 4 the inputs are: 48 number of calls per 60 minutes = 1.92 Erlangs, AHT 144 secs - 85% Answered in 15 secs, Shrinkage 5%, Max Occupancy 95%. and after running the program, it gives the ideal number of CRSs to be 4.

For the same inputs, if the SL is set at 96.3%, the program returns a value of 5.5 so that the ideal number of CRSs becomes 6.

Table 4.5: Optimal number of CSRs for a given SL using Erlang A on a Random day

N	ABS	SL	Occupancy	ASA (s)	% AI	AR
2	2	6.7%	96%	1692.7	6%	33.18%
3	3	63.6%	64%	54.3	59.2%	15.27%
4	4	87.5%	48%	10.7	84.5%	5.93%
$\lceil 5.5 \rceil = 6$	5	96.3%	38.4%	2.4	94.9%	1.95%
$\lceil 6.5 \rceil = 7$	6	99%	32%	0.5	98.5%	0.55%
$\lceil 7.5 \rceil = 8$	7	99.8%	27.4%	0.1	99.6%	0.13%

The ideal number of agents when the service level is set at 85% (which is the industry norm) is 4. The call centre has a maximum of 6 agents and therefore a SL of more than 96.3% is not possible. service level of 99.8% is extremely difficult and might not be realistic in the real world, even though this would be the most desirable outcome and result.

Results for Tuesday

The input parameters for Tuesday were 210 calls per 8 hours, AHT Time 144 seconds, 90% answered in 15 secs, Shrinkage 5%, Max Occupancy 95%. and after running the program, the results returned in the 3rd row indicate the ideal number of CRSs to be 5 at a SL of 87.5%. As previously mentioned, the call centre has a maximum of 6 CRSs and therefore one could make use of the maximum number of CRSs and obtain a greater SL of 94.8%.

Table 4.6: Optimal number of CSRs for a given SL using Erlang A on a peak Tuesday

N	ABS	SL	Occupancy	ASA (s)	% AI	AR
$\lceil 3.5 \rceil = 4$	3	75.5%	60%	50.8	62%	20%
$\lceil 4.5 \rceil = 5$	5	87.5%	53%	37	78%	8%
$\lceil 5.5 \rceil = 6$	5	94.8%	34%	1.8	88%	2%
$\lceil 6.5 \rceil = 7$	6	98.3%	30%	0.2	95%	1.2%

4.6.2 Fitting the Erlang C model

An analysis of the call centre data for the peak Tuesday and a random day was repeated with the same inputs as given in the previous section where the Erlang A model was fitted. The R code in Appendix B8 is used to predict the required number of agents or CRSs to be 7 and 6 respectively as indicated in the tables 4.7 and 4.8. It is noted that the model over states the required CRSs as expected and noted in the literature.

Results for a randomly selected day

Table 4.7: Optimal number of CSRs for a given SL using Erlang C on a Random day

N	ABS	SL	Occupancy	ASA (s)	% AI	AR
3	2	6.7%	96%	1692.7	6%	30.73%
$\lceil 5.5 \rceil = 6$	4	77.5%	48%	10.7	84.5%	4.975%
7	5	86.3%	38.4%	2.4	94.9%	1.56%
$\lceil 8.5 \rceil = 9$	6	99%	32%	0.5	98.5%	0.42%
10	7	99.8%	27.4%	0.1	99.6%	0.1%

For this particular Call center, if the assumptions of the Erlang C hold then, the maximum SL that can be realised on the random day is 77.5% with 6 agents.

Results for Tuesday

Table 4.8: Optimal number of CSRs for a given SL using Erlang C on the peak Tuesday

N	ABS	SL	Occupancy	ASA (s)	% AI	AR
3	2	65.6%	54%	59.3	62.1%	15.51%
$\lceil 4.5 \rceil = 5$	3	71%	36%	82	89%	4.24%
$\lceil 5.5 \rceil = 6$	4	85.1%	27%	1.3	97.4%	0.93%
$\lceil 7.5 \rceil = 8$	5	99.6%	21.6%	0.2	84.5%	0.17%
$\lceil 8.5 \rceil = 9$	6	99.9%	18%	0	99.9%	0.03%

For this particular Call center, if the assumptions of the Erlang C hold then, the maximum SL that can be realised on the peak Tuesday is 85.1% with 6 agents.

The Erlang C does not take into consideration abandoned calls and it has no blockage, (i.e everyone who calls will get through to the queue).

The Erlang C formula or model has a weakness in that it tends to overestimate the required number of staff as seen in the outcome. This is a direct result of its flaws and assumptions such as, people staying on hold indefinitely, yet in reality callers will loose patience and abandon the queue.

Chapter 5

Conclusions and recommendations

This chapter contains the conclusions and some recommendations emanating from the study.

5.1 Conclusion

Managing a call center, with all the uncertainties associated with call arrivals, incidences of abandonment and varying call service times is no doubt an arduous task fraught with undesired outcomes in many instances. As noted in Section 3.5, the queuing primitives which include call arrival rate, call service rate, call abandonment etc. are all stochastic in nature and therefore pose a challenge when they need to be inputs in determining the optimal number of agents to be deployed. To this end, the call center manager would need to consider modeling the operations of the call center through the use of the well developed, tried and tested queuing models to improve and optimize the call center's operations to achieve a pre-set service level.

In this work, it was noted that the call arrival process over time was a stationary process. An investigation into the feasibility of modeling the arrival rate using a non-homogeneous Poisson process where the arrival rate was deemed to be piecewise constant was done. The data at hand did not provide evidence to reject the assertion that the arrival rate function is piece-wise

constant which is like in other studies on call centers that are discussed in Chapter 3.

In the analysis of service times data, the usual "naive" assumption that service times follow an exponential distribution was rejected outright. A histogram of log service times appear to be suggesting that the lognormal distribution could be a suitable candidate for modeling service times like the case of call centers also explored in Chapter 3. The statistical tests of hypotheses, however led to a different conclusion regarding the suitability of the lognormal distribution.

Whilst the Erlang C model is a great tool, it has some limitations mentioned earlier which make it a less desirable tool to use in reality. It assumes that no caller abandons the queue which in reality is not possible. It also has a 'no busy signal' where the queue is assumed open indefinitely. Calls are assumed to follow a fixed arrival pattern and all staff are assumed to be readily available to take any and all calls. All of which are simply not practical! This makes the Erlang C model unrealistic and undesirable to the real life call centre that faces all these challenges on a daily basis.

The Erlang C was modelled for completions sake and it is noted that the Erlang A model is more ideal for a real life call centre and the Erlang C falls short due to its unrealistic assumptions.

5.2 Recommendations

The determination made that service times may not follow a lognormal distribution indicate a cause of concern as this outcome would have been desired. To mind comes the use of the Weibull distribution. According to Kundu and Manglick (2004), the rivalry of the log-normal and Weibull distributions in modeling a skewed data set is well-known in literature. The two distributions usually provide similar data fit for moderate sample sizes and determining which of the two provide the more desirable or nearly correct model is usually a challenge mainly because inferences based on the model will often involve tail probabilities.

Another problem that possibly affect the results is the precision of time measurement at the call server. High precision in the measurement and recording of times could possibly lead to different conclusions.

The issue of to what extent training of call center agents can impact on service times is another aspect that may need probing.

It would be extremely beneficial to both the call center and to the caller if the call center could put a mechanism in place that automatically returned a dropped call to the lost position in the queue or at least return the lost caller to the agent that was attending to the caller prior to the drop taking place.

Overall, the literature has shown that the call center is the nerve centre of a company and is the first point of call for a customer. This service needs to be made available and the agents should at all times be kept well skilled and professional. The call center agents therefore have to always be trained and kept up to date with all company changes and any new products that the caller may require information on. It is highly recommended therefore that an appropriate and professional service be offered to callers who request service. Very few callers should be turned away or told that they cannot get assistance due to a call center's inadequacies. For this to happen, no doubt, the use of mathematical statistics models is an imperative!

Appendix A:

Table A1: Data sample collected from call center

	B	C	D	E	F	G	H	K	L	M	N	O	P	Q	R	S	T
1	Router Date/Time	Router Date	Router Date (dayOfWeek)	Router Time (hh:mm:ss)	Router Time (hh)	Router Time (mm)	Router Time (ss)	Caller	Starter	Router	Response Time (hh:mm:ss	Response Time (ss)	Response Time (mm)	Agent	Call Duration	Call Duration (ss)	Transfer Destination
119	03/01/2017 11:35:58	03/01/2017	Tue	11:35:58	11	35	58	076*******	Call Centre	Call Centre	00:00:08	8	0	2	00:00:38	38	53729
331	04/01/2017 13:33:03	04/01/2017	Wed	13:33:03	13	33	3	083*******	Call Centre	Call Centre	00:00:05	5	0	4	00:01:03	63	
618	05/01/2017 14:16:14	05/01/2017	Thu	14:16:14	14	16	14	061*******	Call Centre	Call Centre	00:00:36	36	1	1	00:00:39	39	
700	05/01/2017 15:27:17	05/01/2017	Thu	15:27:17	15	27	17	081*******	Student Call	Student Call	00:00:08	8	0	5	00:00:34	34	53729
1517	10/01/2017 12:29:59	10/01/2017	Tue	12:29:59	12	29	59	012*******	Student Call	Student Call	00:00:02	2	0	2	00:00:48	48	
1521	10/01/2017 12:31:55	10/01/2017	Tue	12:31:55	12	31	55	076*******	Student Call	Student Call	00:00:05	5	0	5	00:01:31	91	53736
1811	11/01/2017 11:06:10	11/01/2017	Wed	11:06:10	11	6	10	011*******	Student Call	Student Call	00:00:07	7	0	6	00:01:40	100	58201
1848	11/01/2017 11:38:35	11/01/2017	Wed	11:38:35	11	38	35	062*******	Student Call	Student Call	00:00:57	57	1	4	00:09:58	598	
3575	25/01/2017 09:21:18	25/01/2017	Wed	09:21:18	9	21	18	031*******	Student Call	Student Call	00:00:04	4	0	2	00:00:28	28	53729
3578	25/01/2017 09:31:09	25/01/2017	Wed	09:31:09	9	31	9	084*******	Student Call	Student Call	00:00:02	2	0	1	00:00:53	53	
3599	25/01/2017 10:47:53	25/01/2017	Wed	10:47:53	10	47	53	083*******	Student Call	Student Call	00:00:03	3	0	4	00:00:53	53	
3600	25/01/2017 10:51:21	25/01/2017	Wed	10:51:21	10	51	21	011*******	Student Call	Student Call	00:00:08	8	0	3	00:01:51	111	
3605	25/01/2017 11:07:35	25/01/2017	Wed	11:07:35	11	7	35	045*******	Student Call	Student Call	00:00:05	5	0	6	00:01:40	100	50521
3607	25/01/2017 11:20:43	25/01/2017	Wed	11:20:43	11	20	43	012*******	Student Call	Student Call	00:00:05	5	0	5	00:00:57	57	50184
3619	25/01/2017 12:04:45	25/01/2017	Wed	12:04:45	12	4	45	081*******	Student Call	Student Call	00:00:04	4	0	3	00:01:35	95	53729
3625	25/01/2017 12:34:57	25/01/2017	Wed	12:34:57	12	34	57	012*******	Student Call	Student Call	00:00:04	4	0	2	00:00:11	11	
3629	25/01/2017 12:49:15	25/01/2017	Wed	12:49:15	12	49	15	086*******	Student Call	Student Call	00:00:06	6	0	5	00:01:56	116	
3647	25/01/2017 13:39:44	25/01/2017	Wed	13:39:44	13	39	44	067*******	Student Call	Student Call	00:00:02	2	0	1	00:00:49	49	58201
3649	25/01/2017 13:43:00	25/01/2017	Wed	13:43:00	13	43	0	011*******	Student Call	Student Call	00:00:03	3	0	6	00:00:57	57	
3676	25/01/2017 15:29:44	25/01/2017	Wed	15:29:44	15	29	44	012*******	Student Call	Student Call	00:00:02	2	0	5	00:00:31	31	53729
3678	25/01/2017 15:38:47	25/01/2017	Wed	15:38:47	15	38	47	083*******	Student Call	Student Call	00:00:03	3	0	2	00:00:33	33	
3681	25/01/2017 15:50:16	25/01/2017	Wed	15:50:16	15	50	16	061*******	Student Call	Student Call	00:00:06	6	0	2	00:01:53	113	50184
3802	27/01/2017 08:16:25	27/01/2017	Fri	08:16:25	8	16	25	078*******	Student Call	Student Call	00:00:03	3	0	5	00:02:12	132	
3815	27/01/2017 09:20:20	27/01/2017	Fri	09:20:20	9	20	20	083*******	Student Call	Student Call	00:00:03	3	0	3	00:00:33	33	53729
3818	27/01/2017 09:31:47	27/01/2017	Fri	09:31:47	9	31	47	082*******	Student Call	Student Call	00:00:02	2	0	6	00:00:16	16	52843

Appendix B:

Appendix B1: R code for Augmented Dickey-Fuller Test

```
> library("tseries", lib.loc="~/R/win-library/3.5")

'tseries' version: 0.10-46

'tseries' is a package for time series analysis and computational finance.

See 'library(help="tseries")' for details.

Warning message:
package 'tseries' was built under R version 3.5.3

grp_data <- df %% group_by(Router.Date, Router.Time.Hour)
%% summarise(Frq = n())

adf.test(grp_data\$Frq, alternative = "stationary")


data:  grp_data\$Frq

alternative hypothesis: stationary
Warning message: In adf.test(electricity) : p-value smaller than
printed p-value

Warning message:
In adf.test(grp_data\$Frq, alternative = "stationary") :
p-value smaller than printed p-value
```

Appendix B2: R code for distribution of calls over time

```
ggplot(df,aes(x = factor(Router.Time..hh.))) +
+ geom_bar(aes(y=(..count..)/sum(..count..)), stat="count",
width=0.7,fill="steelblue") + + coord_flip() +
```

```
+ geom_text(aes(y=..count../sum(..count..),
label=paste0(..count../sum(..count..)*100,"%")), hjust=1, stat="count") +
xlab("Hours of day") + ylab("Number of calls / hour")
```

Appendix B3: R code for distribution of arrivals calls per day during the week

```
ggplot(df,aes(x = factor(Router.Date..dayOfWeek.))) +
+ geom_bar(aes(y=(..count..)/sum(..count..)),stat="count",width=0.7,
fill="steelblue") + + coord_flip() +  + geom_text(aes(y=..
count../sum(..count..),
label=paste0(..count../sum(..count..)*100,"%")), hjust=1, stat="count") +
xlab("Days of the week") + ylab("Number of calls"))
```

Appendix B4: R code for distribution of calls on 1st day in March 2019

```
one_day_dat <- filter(df, Router.Date == '01/03/2017')
table(one_day_dat\$Router.Time..hh.)
```

Appendix B5: R code for Log Service Times

```
ggplot(data, aes(x=LogDuration)) + geom_histogram(aes(y=stat(density)),
color = 'blue', fill='white') + labs(x='Log(Service Time)',
y='Proportion') + stat_function(fun = dnorm, args
= list(mean=mean(dat$LogDuration),
sd=sd(dat$LogDuration)), lwd=2)
```

Appendix B6: R code for customer caller patience times

```
library(survival)
ds = data.frame(data1\$Call.Duration.Second, data1\$surv)
fit = survfit(Surv(data1.Call.Duration.Second, data1.surv)
~ data1\$surv, data=ds)
autoplot(fit) + xlab('Time to Failure (in seconds)') +
ylab('Proportion of Survivors')
```

Appendix B7: R code for Survival Function of Virtual Waiting Time

```
library(survival)
ds = data.frame(data1\$Response.Time.Second, data1\$surv)
fit = survfit(Surv(data1.Response.Time.Second,data1.surv) ~ 1,
data=ds)
summary(fit)\$table
summary(fit)
#plotting the survival function of the waiting time
>autoplot(fit) + xlab('Time to Failure (in seconds)') +
ylab('Survival Proportion')
```

Appendix B8: R code for Determining the Required Number of Staff

```
library(tidyverse)
RequiredAgents <- function(rate, duration, target, SL_target, interval = 60) {
agents <-round(intensity(rate, duration, interval) + 1) # input is fixed at 6 + 1
ServiceLevel <- service_level(agents, rate, duration, target, interval) #random da
while (ServiceLevel < SL_target * (SL_target > 1) / 100) {
agents <- agents + 1
ServiceLevel <- service_level(agents, rate, duration, target, interval)
}
return(c(agents, ServiceLevel))
}
```

References

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM.

Anderson, T. W. (2011). *Anderson–Darling Tests of Goodness-of-Fit*, pages 52–54. Springer Berlin Heidelberg, Berlin, Heidelberg.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769.

Angus, I. (2001). An introduction to erlang b and erlang c. *Telemanagement*, 187:6–8.

Avramidis, A. N., Deslauriers, A., and L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50.

Cruz, F. R., Smith, J. M., and Medeiros, R. (2005). An m/g/c/c state-dependent network simulation model. *Computers & Operations Research*, 32(4):919–941.

Fang, Y. (2001). Hyper-erlang distribution model and its application in wireless mobile networks. *Wireless Networks*, 7(3):211–219.

Gail, M. and Gastwirth, J. (1978). A scale-free goodness-of-fit test for the exponential distribution based on the gini statistic. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):350–357.

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

Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318.

Knessl, C. and van Leeuwaarden, J. S. (2015). Transient analysis of the erlang a model. *Mathematical Methods of Operations Research*, 82(2):143–173.

Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1):41–59.

Kundu, D. and Manglick, A. (2004). Discriminating between the weibull and log-normal distributions. *Naval Research Logistics (NRL)*, 51(6):893–905.

Liao, S., Koole, G., Van Delft, C., and Jouini, O. (2012). Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR spectrum*, 34(3):691–721.

Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402.

Mandelbaum, A., Sakov, A., and Zeltyn, S. (2000). Empirical analysis of a call center. *URL http://iew3. technion. ac. il/serveng/References/ccdata. pdf. Technical Report*.

Mandelbaum, A. and Zeltyn, S. (2007). Service engineering in action: the palm/erlang-a queue, with applications to call centers. *Advances in services innovations*, pages 17–45.

Palm, C. (1953). Methods of judging the annoyance caused by congestion. *Tele*, 4(189208):4–5.

Pearson, K. (1916). Ix. mathematical contributions to the theory of evolution.—xix. second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 216(538-548):429–457.

Qiao, S. and Qiao, L. (1998). A robust and efficient algorithm for evaluating erlang b formula.

Robbins, T. R., Medeiros, D. J., and Dum, P. (2006). Evaluating arrival rate uncertainty in call centers. In *Proceedings of the 38th conference on Winter simulation*, pages 2180–2187. Winter Simulation Conference.

Robbins, T. R., Medeiros, D. J., and Harrison, T. P. (2010). Does the erlang c model fit in real call centers? In *Simulation Conference (WSC), Proceedings of the 2010 Winter*, pages 2853–2864. IEEE.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Tunnicliffe, G., Murch, A., Sathyendran, A., and Smith, P. (1998). Analysis of traffic distribution in cellular networks. In *VTC'98. 48th IEEE Vehicular Technology Conference. Pathway to Global Wireless Revolution (Cat. No. 98CH36151)*, volume 3, pages 1984–1988. IEEE.

Whitt, W. (1999). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24(5):205–212.

Yom-Tov, G. B. and Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299.

Zhan, D. and Ward, A. R. (2013). Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management*, 16(2):220–237.