Dissertação

Mestrado em Gestão de Sistemas de Informação Médica

Caracterização de uma Fila de Espera de um Serviço Hospitalar – Um Estudo de Caso

Mário João Dias Carvalho

Leiria, junho de 2015

Dissertação

Mestrado em Gestão de Sistemas de Informação Médica

Caracterização de uma Fila de Espera de um Serviço Hospitalar – Um Estudo de Caso

Mário João Dias Carvalho

Dissertação de Mestrado realizada sob a orientação do Doutor Rui Filipe Vargas de Sousa Santos, Professor da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e coorientação da Doutora Liliana Catarina Rosa Ferreira, Professora da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria.

Leiria, junho de 2015

Dedicatória

À Minha Família

Agradecimentos

Os meus sinceros agradecimentos aos meus orientadores, Doutor Rui Santos e Doutora Liliana Ferreira, por todo o apoio e aprendizagem que me proporcionaram. Sempre disponíveis a ajudar-me e a orientar-me da melhor maneira, sem eles sentir-me-ia perdido neste mundo tão vasto como são as áreas da estatística e da programação.

Agradeço também ao Hospital de Santo André de Leiria por ter fornecido os dados sem os quais não seria possível realizar este estudo.

Resumo

Este estudo tem como principal objetivo caracterizar as filas de espera do serviço de urgência de uma unidade de saúde, mais concretamente do Hospital de Santo André (HSA) que pertence ao Centro Hospitalar de Leiria.

A caracterização foi feita recorrendo aos dados sobre o percurso dos utentes, durante o ano de 2014, no que diz respeito à hora de chegada ao hospital, hora de admissão, hora de triagem, classificação do utente na triagem, hora de consulta, hora de saída, bem como outras medidas e/ou características que influenciam o tempo que um utente permanece nas urgências.

Neste sentido, inicialmente fez-se um estudo sobre a teoria das filas de espera e sobre os diversos modelos existentes. Considerando o número de servidores, as distribuições estatísticas dos tempos de chegada e tempos de espera, a disciplina das filas e as várias fases do sistema, adequou-se o modelo apropriado ao caso de estudo.

Implementando a simulação, através do *software* R, criou-se uma matriz com os dados dos utentes comparando-os com os dados da base de dados fornecida. Parte essencial da simulação foi a componente das prioridades que, através do sistema de Triagem de Manchester, classifica os utentes segundo a sua urgência.

Esta simulação permitiu criar vários cenários e, assim, obter conclusões de modo a contribuir para um melhor conhecimento do comportamento do sistema e sobre a eficiência das filas de espera do serviço de urgência do HSA.

Deste modo, o principal objetivo deste estudo será fornecer informação relevante que permita investigar eventuais melhoramentos de eficiência no funcionamento do serviço, aumentando, por outro lado, a satisfação dos utentes, no sentido que permitirá encontrar o equilíbrio entre o custo associado à prestação do serviço e o custo inerente à espera até obter esse serviço.

Palavras-chave: *Modelos de filas de espera; Distribuições estatísticas; Software R; Simulação; Serviço de urgências; Triagem.*

Abstract

This study aims to characterize the waiting lines of the emergency unit of a health care facility, the "Hospital de Santo André (HSA)" which belongs to the Hospital of Leiria.

The characterization was done using the hospital's patient data during 2014, taking into account the following steps: time of arrival at the hospital, admission time, patients triage time, patients triage classification, time patient was seen by a doctor, time patient left the hospital, as well as other factors that influenced the time a patient remained in the emergency unit.

Initially a study was done about the theory of the existing waiting lines models. Then, considering the number of servers, the statistical distribution of arrival and waiting times, the discipline of the queues and the several phases of the system, an appropriate model was adapted for this case study.

By performing the simulation, using the R software, a matrix with the patients data was created and compared with the data provided by the database. An essential part of this simulation was the priority component which, through the triage system of Manchester, classifies the patients by level of urgency.

This simulation allows the creation of various scenarios and to obtain conclusions that contribute to a better understanding of the systems behavior and efficiency of the waiting queues of the emergency services of the HSA.

Thus, the main goal of this study is to provide relevant information that would enable to investigate potential future efficiency improvements in the operation of the service, increasing, on the other hand, the satisfaction of the patients, and allow for a balance between the cost associated in providing the service and the cost of the waiting period before receiving that service.

Keywords: *Queueing models; Statistical distributions; R software; Simulation; Emergency service; Triage.*

Lista de figuras

Lista de tabelas

Lista de siglas

ATS	Australasian Triage Scale
CTAS	Canadian Triage and Acuity Scale
ECDF	Empirical Cumulative Distribution Function
ESI	Emergency Severity Index
FCFS	First Come First Served
FDP	Função Densidade de Probabilidade
FIFO	First In First Out
GD	Disciplina Geral
HSA	Hospital de Santo André
IID	Independente e Identicamente Distribuídos
LCFS	Last Come First Served
LIFO	Last In First Out
METTS	Medical Emergency Triage and Treatment System
MTS	Manchester Triage Scale
NA	Not Available
NPRP	Nonpreemptive Priority Models
ns	Número de servidores na admissão
nsc	Número de servidores nas consultas
nst	Número de servidores na triagem
PRP	Preemptive Priority Models
SIRO	Service In Random Order
Tc	Taxa média de chegada dos clientes à admissão
tea	Média do tempo de espera para a admissão
tec	Média do tempo de espera para a consulta
tet	Média do tempo de espera para a triagem
Ts	Taxa média de serviço de um servidor da admissão
tsa	Média do tempo de serviço da admissão
Tsc	Taxa média de serviço de um servidor da consulta
tsc	Média do tempo de serviço da consulta
Tst	Taxa média de serviço de um servidor da triagem
tst	Média do tempo de serviço da triagem

Índice

1. Introdução

Com um papel cada vez mais importante no Sistema Nacional de Saúde surgem os serviços de urgência, onde os tempos de espera são um fator importante na satisfação dos utentes.

O serviço de urgência existe de modo a proporcionar aos utentes um atendimento rápido em situações de potencial risco para a saúde. Uma urgência é qualquer situação cuja demora de diagnóstico ou de tratamento apresente grave risco ou prejuízo para a vítima, como, por exemplo, traumatismos graves, intoxicações agudas, queimaduras, crises cardíacas ou respiratórias. Os utentes em situações graves necessitam de atendimento mais rápido em relação aos menos graves, visto que uma espera prolongada pode comprometer o seu estado de saúde. É então necessário que exista um método que permita classificar a prioridade clínica dos pacientes, por exemplo, o Sistema de Triagem de Manchester.

Após efetuar o registo de entrada no serviço de urgência o utente é encaminhado para um gabinete de triagem, onde é submetido a uma observação prévia, com identificação de um conjunto de sintomas ou sinais que permitem atribuir um grau de prioridade clínica no atendimento, bem como um tempo de espera máximo recomendado até à primeira observação médica.

Num serviço de urgência, um utente encontra-se sujeito a tempos de espera, sendo estes constantemente motivos de reclamações devido às esperas prolongadas. Um serviço de urgência em que os utentes são atendidos rapidamente é considerado um bom serviço de urgência [5].

O objetivo principal da teoria das filas de espera é otimizar o desempenho de um sistema, de modo a reduzir os seus custos operacionais e a aumentar a satisfação do utente [17]. O conhecimento de alguns parâmetros no estudo de filas de espera, tais como o tempo médio de chegadas de utentes, o tempo de espera até receberem assistência, o número de utentes que se encontram em simultâneo à espera de serem atendidos e a taxa de ocupação de cada recurso, permite, por um lado, quando os recursos são limitados, determinar qual o conjunto de regras de prioridade que maximizam as taxas de atendimento e, por outro, dimensionar os recursos de modo a conter os tempos de espera e o número de utentes dentro de limites máximos aceitáveis.

A solução de problemas desta natureza é normalmente encontrada procurando, através da simulação, o melhor conjunto de regras e a melhor dimensão dos recursos. A simulação pode desempenhar um papel muito importante neste contexto. Em vez de procurar avaliar diretamente a performance do serviço de urgência, pode-se simulá-lo, utilizando distribuições de probabilidade que permitem gerar aleatoriamente vários eventos que ocorrem nas diversas unidades que o integram (admissão, triagem e consulta) [9, p. 1085]. Como se compreende, seria muito difícil realizar experiências diretamente num sistema de serviço de urgências. Aumentar ou diminuir os funcionários de serviço, pondo em risco a própria saúde dos utentes, ou alterar os tempos de serviço de modo a analisar o comportamento do sistema não é, de todo, viável. Então, uma forma de contornar este problema, passa pela construção de um modelo baseado nas principais características de um serviço de urgência e de realizar experiências sobre ele, registando os resultados da simulação.

Recorrendo à simulação de filas de espera, consegue-se demonstrar o comportamento típico das filas de espera de um serviço de urgências e esclarecer os responsáveis clínicos sobre as variáveis que o influenciam.

Este documento está estruturado em 5 capítulos. No capítulo 2, é apresentado o estado da arte com uma revisão da literatura especializada da área, procurando analisar, de um modo geral, os diversos modelos das filas de espera, conciliando-os com a realidade existente nas unidades de saúde. No capítulo 3, explica-se a metodologia utilizada, com especial incidência nas características da simulação implementada através do *software* R. São também comparadas as distribuições estatísticas entre o modelo de simulação e a base de dados real, de modo a verificar se as duas amostras seguem a mesma distribuição. Procurar-se, ainda, encontrar os valores mais adequados para os parâmetros da simulação de modo a ficar com um modelo minimamente aceitável, tendo em conta a comparação que se faz com a base de dados real. No capítulo 4, criam-se diversos cenários, alterando valores dos parâmetros do modelo, permitindo, assim, testar a eficiência do sistema. Será também testada a estabilidade do modelo e apresentam-se os resultados e as respetivas análises. Por fim, no capítulo 5, apresentam-se as conclusões, reflexões finais e o trabalho a desenvolver no futuro.

2. Revisão da literatura

Neste capítulo, apresenta-se uma revisão bibliográfica com o intuito de clarificar as ideias relacionadas com o tema principal desta dissertação, as Filas de Espera, procurando analisar os diversos estudos e, sempre que possível, conciliar a teoria com a realidade existente nas unidades de saúde. Assim, começa-se por referir um pouco da história das filas de espera, seguindo-se uma análise mais profunda acerca da teoria das filas de espera, com especial incidência nos modelos utilizados em diversos sistemas.

2.1. História das Filas de Espera

Uma fila de espera apresenta um comportamento dinâmico e mal compreendido pela maioria das pessoas. Este comportamento foi objeto de estudo e teorização por parte de matemáticos ao longo do século XX, sobretudo a partir dos anos 40. As leis matemáticas descritas eram contudo muito complexas, o que tornou esta matéria tratável apenas por especialistas. Nos anos 80-90 com o desenvolvimento das técnicas de simulação, foi possível descrever de forma muito mais simples e compreensível o comportamento das filas de espera.

Agner Krarkup Erlang (Janeiro 1, 1878-Fevereiro 3, 1929) foi o matemático e engenheiro dinamarquês que idealizou pela primeira vez os conceitos de Engenharia de Tráfego e da Teoria das Filas. A teoria das filas foi desenvolvida para fornecer modelos matemáticos que prenunciam o comportamento de diversos sistemas que tentam fornecer um atendimento adequado às necessidades dos clientes.

Foi quando Erlang trabalhava na empresa "*Copenhagen Telephone Company*", em 1904, que teve que resolver um clássico problema de determinar quantos circuitos são necessários para fornecer um atendimento aceitável nas chamadas telefónicas. O seu raciocínio ajudou-o a perceber que a matemática resolveria outro problema, que seria, quantos operadores de telefone são necessários para atender um número de chamadas telefónicas determinadas previamente. Nessa época, a maioria das centrais telefónicas usava trabalhadores como operadores para gerir as chamadas telefónicas, que ligavam os

fios telefónicos nas tomadas elétricas das placas com circuitos. Com o seu trabalho, Erlang pretendia ajudar a determinar os requisitos de capacidade que o sistema telefónico deveria ter para assegurar um nível de serviço adequado à procura. O problema colocava-se porque essa capacidade do sistema não deveria ser tão grande ao ponto de originar muita ociosidade no sistema, nem tão pequena ao ponto de originar constantes congestionamentos, considerando-se ainda todo o investimento financeiro envolvido.

Erlang trabalhou no desenvolvimento da área de tráfego nos sistemas de chamadas telefónicas e publicou o seguinte [1]:

- Em 1909, "*The Theory of Probabilities and Telephone Conversations*" onde provou que a distribuição de Poisson se aplica ao tráfego aleatório de chamadas telefónicas.
- Em 1917, "Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges" que inclui a sua fórmula clássica de tempo de espera e tempo perdido.

Existiam avanços nas aplicações telefónicas mas na teoria das filas não houve um avanço semelhante. Foi no fim da 2.ª Guerra Mundial e na década dos anos 50 que os estudos de Erlang foram aplicados em problemas mais gerais, incidindo também na aplicação de filas de espera em negócios, fazendo com que as aplicações em áreas para além dos sistemas de telefone começassem a evoluir. Posteriormente, a Teoria das Filas de Espera foi aplicada em múltiplos setores de atividade, como a engenharia de produção, a banca, a exploração de linhas aéreas e as emergências hospitalares.

2.2. Teoria das Filas de Espera

Gerir eficazmente um serviço de urgência de cuidados de saúde, o qual se caracteriza pelo desconhecimento do momento exato em que um paciente vai aparecer e do tipo e quantidade de recursos que irá necessitar, passa pelo conhecimento e compreensão dos fenómenos de filas de espera.

Segundo um estudo coordenado por Cabral, M. *et al.* (2002) [4], as "demoras" são o principal motivo (em igualdade com a "má assistência/erro médico") de reclamações dos

utentes contra os serviços públicos de saúde, em Portugal. No mesmo estudo, o tempo de espera por uma consulta (no centro de saúde) e o tempo de espera antes de ser atendido foi considerado por 43% dos utentes inquiridos como sendo mau ou muito mau.

A espera por um serviço ocorre sempre que o número de pessoas que pretendem um serviço excede o número de pessoas que o sistema consegue atender de imediato. Organizações prestadoras de serviços lidam constantemente com filas de espera, portanto, esta é uma fonte de preocupação importante na qualidade do serviço que prestam.

Uma gestão de filas eficiente tem como um dos seus objetivos principais a redução do congestionamento. Um Serviço de Saúde que se preocupe com o tempo de espera a que os utentes estão sujeitos dará, certamente, um passo em frente na melhoria da qualidade dos serviços que presta.

Existem filas de espera nas urgências, na marcação de consultas de especialidade (nos hospitais), na marcação de consultas de medicina geral e familiar (nos centros de saúde), na marcação de intervenções cirúrgicas, nos centros de atendimento de urgência, nos centros de vacinação, enfim, num inúmero conjunto de situações. E porquê a existência de longas filas de espera nestes casos? O problema central das filas de espera deriva do *trade-off* (situação em que há conflito de escolha) entre o custo de prestar um serviço mais rápido e o custo da espera (ver Figura 2.1).

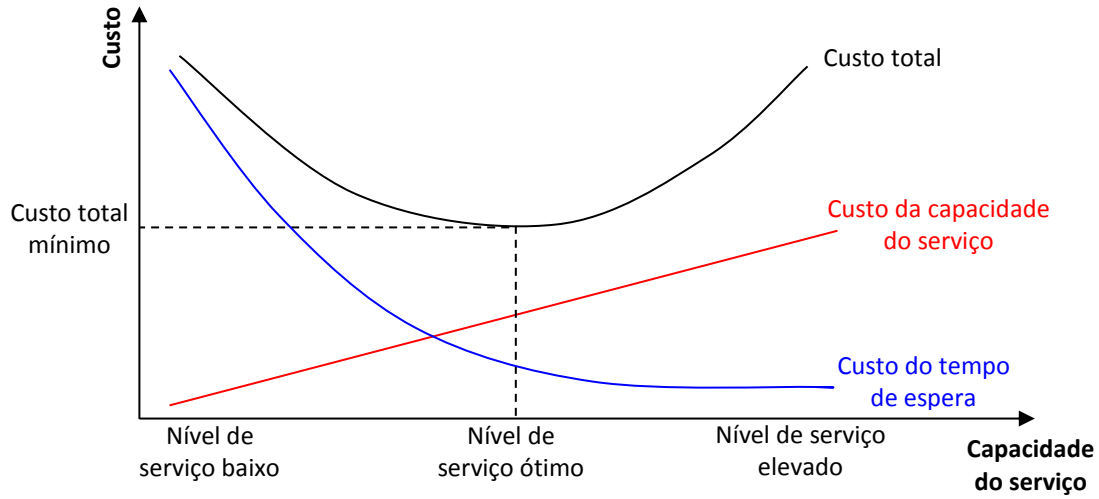


Figura 2.1 - *Trade-off* entre o custo da capacidade do serviço e o custo do tempo de espera

As filas de espera existem devido, em grande parte, à insuficiente capacidade do serviço que, por sua vez, existe devido ao custo elevado a considerar pelo aumento dessa mesma

capacidade. Por outro lado, o *trade-off* representado na Figura 2.1, adaptada de [9, p. 912], nem sempre é considerado, sendo apenas avaliado o custo do serviço isoladamente, o que se repercute numa capacidade de serviço diminuta, criando filas de espera.

Analisando a Figura 2.1, com a capacidade de serviço mínima, o custo do tempo de espera está no máximo. À medida que a capacidade de serviço vai aumentando, existe uma redução do número de utentes na fila e do seu tempo de espera, o que diminui o custo do tempo de espera. Em situações de *trade-off* é necessário encontrar o ponto de equilíbrio, ou seja, qual a capacidade do serviço que minimiza o custo total e que corresponde ao somatório do custo da capacidade do serviço com o custo do tempo de espera.

Enquanto que o custo por acrescentar mais capacidade ao sistema é relativamente simples de quantificar (mais recursos humanos, mais equipamento, mais instalações, entre outros), o custo do tempo de espera é muito difícil de quantificar. Como quantificar o tempo perdido por um utente numa fila de espera para marcação de uma consulta com o seu médico de família? Como quantificar o tempo que um utente fica à espera de uma ambulância, quando a sua vida está em risco? Como quantificar o tempo que um utente fica à espera para ser observado nas urgências, enquanto o seu estado de saúde se pode degradar?

Dada a dificuldade de quantificar o custo da espera, a resolução do problema das filas de espera com base no custo total não é exequível.

Neste capítulo serão, assim, abordados modelos quantitativos para a gestão de filas de espera, que indicam o desempenho esperado de um sistema de filas de espera, não sendo necessário quantificar o custo da espera.

Para além destes modelos, existe um conjunto de sugestões que pode ser bastante útil na gestão de filas de espera [6] e que também não poderia deixar de ser contemplado neste capítulo:

- Determinar o tempo de espera aceitável pelos utentes e estabelecer objetivos operacionais de acordo com esse tempo;

- Distrair a atenção dos utentes enquanto esperam. Música, televisão, vídeo, revistas ou outro tipo de entretenimento pode ajudar os utentes a distraírem-se do facto de estarem à espera;

- Informar os utentes do tempo de espera previsto. Este aspeto é particularmente importante quando o tempo de espera é mais longo do que o normal. É desejável informar o utente do porquê desta situação e de quais os esforços que estão a ser efetuados para resolver essa questão;

- Os colaboradores que não estão no atendimento devem estar fora do campo visual dos utentes. Os utentes, ao constatarem que existem mais recursos em determinado serviço que poderiam potencialmente estar no atendimento, mas que estão a realizar outras atividades, ficam ainda mais insatisfeitos com o tempo de espera, pois pensam que este poderia ser facilmente reduzido;

- Segmentar os utentes. Se existirem grupos de situações/utentes cujo tempo de atendimento é semelhante e é notoriamente mais rápido do que outro(s) grupo(s) de situações/utentes, então esse grupo deve ter uma fila dedicada, para que não tenham de esperar por utentes mais morosos;

- Formar os colaboradores para serem simpáticos. A simpatia de quem serve o utente pode ajudar a atenuar/ultrapassar o sentimento negativo do utente provocado por uma longa espera;

- Encorajar os utentes a solicitar o serviço em períodos de tempos de procura baixa. Informar os utentes sobre quais os períodos do dia com menos e com mais afluência. Esta medida pode atenuar as variações acentuadas de afluência de utentes ao longo do dia em alguns serviços.

A aplicação dos modelos analíticos em simultâneo com a aplicação de algumas (ou todas) das sugestões indicadas permite, com certeza, uma gestão mais eficaz do problema das filas de espera.

2.2.1 O sistema de filas de espera

Basicamente, um modelo de fila de espera representa um sistema de serviço onde um cliente se dirige a um ou mais servidores para ser atendido [8]. Se houver um servidor livre, o cliente poderá ser atendido de imediato, mas se todos estiverem ocupados terá de esperar numa fila pelo atendimento.

A Figura 2.2 representa um sistema básico de filas de espera com alguns dos seus elementos principais: a chegada de clientes ao sistema, a fila de espera e o servidor (quem presta o serviço). A disciplina da fila, a capacidade do sistema, o número de servidores e o número de fases que compõem o sistema são outras das características importantes de um sistema de filas de espera [7].

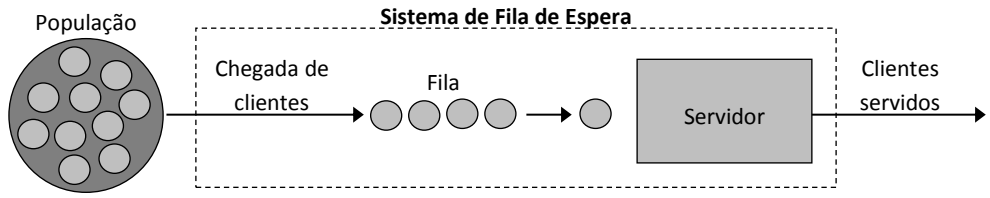


Figura 2.2 - Representação básica do sistema de filas de espera

Os clientes que requerem um determinado serviço chegam de uma determinada população e ao entrarem no sistema de filas de espera juntam-se à fila. Cada membro da fila de espera é selecionado para ser servido de acordo com uma regra, denominada por disciplina da fila. O serviço requerido é executado pelo servidor, após o qual o cliente sai do sistema de filas de espera.

Gerir um sistema de filas de espera é um processo complexo devido, entre outros motivos, à incerteza associada à chegada dos clientes ao sistema (quantos clientes vão chegar e a que ritmo) e ao servidor (tempo que demora a prestar o serviço).

Sendo as "Chegadas dos clientes" e o "Tempo do servidor" duas variáveis, em princípio, aleatórias, é necessário investigar qual a distribuição estatística que essas variáveis seguem. Este é o ponto de partida para a utilização de modelos analíticos ou técnicas de simulação para auxiliar a gestão de filas de espera.

De seguida, serão aprofundadas as características principais de um sistema de filas de espera. A abordagem será feita em termos gerais, mas quando se fizer referência ao caso específico da urgência hospitalar a palavra "clientes" será substituída pela palavra "utentes".

2.2.1.1 População ou Fonte

Os clientes que vão chegar ao sistema são provenientes de uma dada população. Uma das características da fonte ou população é o seu tamanho que pode ser finito ou infinito, ou seja, se existe limite ou não de número de potenciais clientes.

Uma população infinita significa que não existem restrições de chegada, isto é, quando a probabilidade de ocorrer uma nova chegada não é influenciada pelo número de clientes que já se encontram no sistema, o que poderá implicar que a chegada de clientes pode ultrapassar a capacidade do sistema, a qualquer altura.

Uma fonte é finita quando o número de clientes admitidos no sistema é limitado. Neste caso, o modelo analítico é mais complicado, pois o número de clientes dentro do sistema (na fila ou a ser servidos) afeta o número de clientes fora do sistema. Este modelo deve ser adotado sempre que o ritmo a que os clientes são gerados pela fonte depende significativamente do número de clientes que estão dentro do sistema [17].

A análise do sistema torna-se geralmente mais simples quando se considera uma população infinita. Na prática, as situações em que há um conjunto muito numeroso de clientes potenciais podem ser modelados como tal. Um exemplo de uma população infinita, na prática, é a chegada dos utentes ao serviço de urgências de uma unidade hospitalar. Por uma questão de simplificação de cálculos, na maioria dos casos, considera-se a população infinita, exceto naqueles em que o número de clientes que pode chegar ao sistema, num dado intervalo de tempo, depende significativamente do tamanho da população nesse intervalo. Por exemplo, numa situação com 10 máquinas, a probabilidade de se avariar uma máquina na próxima hora depende do número de máquinas que estejam a funcionar.

2.2.1.2 Chegadas

Para modelar um sistema de filas de espera é necessário caracterizar a chegada dos clientes ao sistema. A dimensão da chegada pode ser unitária, quando os clientes chegam um a um, ou em grupo. Neste caso, deve ser determinada a distribuição de probabilidade que descreve o tamanho do grupo [7]. O processo de chegada é a descrição de como os clientes procuram o serviço. O padrão das chegadas pode ser descrito pelo tempo entre duas chegadas consecutivas (tempo entre chegadas) ou pelo número de chegadas por unidade de tempo (distribuição das chegadas) [1]. Pode ser controlável, por exemplo quando existem inscrições em dias fixos, ou incontrolável, como é o caso da chegada de utentes ao serviço de urgência de um hospital. O tempo decorrido entre chegadas consecutivas é definido por *inter-arrival time* [9, 16].

Se os clientes chegam em intervalos fixos de tempo, o processo de chegadas é dito constante ou determinístico. Por outro lado, as chegadas são aleatórias no tempo quando os

intervalos de tempo entre chegadas sucessivas não podem ser previstos. Neste caso, as chegadas formam um processo estocástico, sendo necessário usar as distribuições de probabilidade. Um processo estocástico é um modelo de probabilidade que descreve a evolução de um sistema aleatório no tempo [13].

Num caso como no outro, interessa sobretudo a taxa de chegada que é o número de clientes que, em média, chega ao sistema por unidade de tempo, tal como ilustrado no exemplo da Figura 2.3, adaptada de [17, p. 352]. Esta taxa, habitualmente denotada por λ, pode ser independente do número de clientes existente no sistema ou dependente deste. Neste último caso, se o número de clientes for, por exemplo, n, escreve-se λ_n.

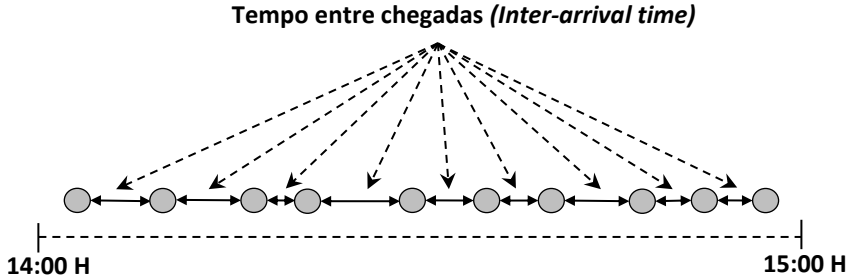


Figura 2.3 - Medidas de padrões de chegadas

No exemplo da Figura 2.3, durante uma hora (14:00-15:00), houve 10 chegadas. Para retirar conclusões mais fiáveis seria necessário obter mais dados, registados durante mais dias e, preferencialmente, no mesmo intervalo de tempo. Considerando que estes dados se verificam também com outros registos, então, a taxa média de chegada é de 10 clientes por hora. Como o espaço existente entre as várias chegadas (tempo entre chegadas ou *inter-arrival time*) não é uniforme, provavelmente será aceitável considerar-se que se está na presença de uma distribuição exponencial negativa, sendo comum designar esta distribuição apenas por distribuição exponencial, termo que se usará daqui em diante.

Se dividirmos o número médio de clientes que chegaram naquele período de tempo, que neste caso foi de 60 minutos, obtém-se a média, por exemplo, da distribuição exponencial que é o tempo médio entre as chegadas. Então, o tempo médio entre chegadas será 60/10 = 6 minutos. Concluindo o exemplo, a taxa média de chegadas pode ser convertida em tempo médio entre chegadas, ou seja, a taxa média de chegadas de 10 clientes/hora corresponde a um tempo entre chegadas de 6 minutos. Por outras palavras, um cliente chega, em média, de 6 em 6 minutos, perfazendo a chegada de 10 clientes por hora.

Por ser mais prática de usar do que a distribuição exponencial, as chegadas podem ser caracterizadas pela distribuição de Poisson. Esta, representada na Figura 2.4 com λ=10, é

uma distribuição discreta que mostra a probabilidade associada a cada número de chegadas durante um determinado período de tempo, onde a média e a variância são iguais [17].

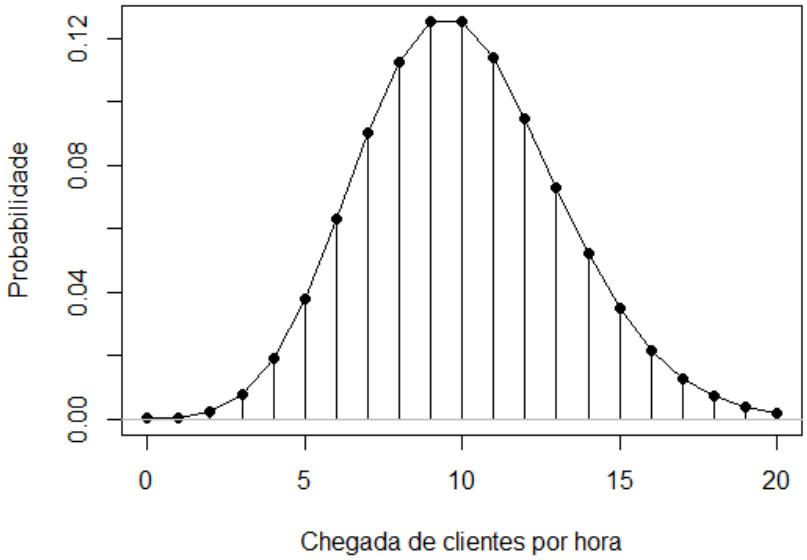


Figura 2.4 - Distribuição de Poisson com λ=10

As taxas de chegada são, normalmente, probabilísticas e os modelos usam, comummente, uma distribuição determinística (quando as chegadas são especificadas num período de tempo ou os tempos entre chegadas são constantes), exponencial (como modelos de distribuição para tempos entre chegadas ou tempos de serviço), Poisson (como distribuição do número de chegadas durante um período específico de tempo), Erlang (como modelos de distribuição de tempos entre chegadas ou tempos de serviço) e variantes destas distribuições [1].

Ao caracterizar um modelo de filas de espera, quanto à probabilidade das chegadas, é comum partir do pressuposto que as chegadas seguem um processo de Poisson. O nome deve-se ao facto de que o número de chegadas num determinado intervalo de tempo seguir, precisamente, uma distribuição de Poisson [8].

Uma característica do processo de Poisson é que o tempo entre chegadas consecutivas (*inter-arrival time*) tem uma distribuição exponencial, sendo ainda caracterizada pelo facto de o tempo da próxima chegada ser independente de quando ocorreu a última. Por outras palavras, se as chegadas seguem o processo de Poisson, o número de chegadas em qualquer intervalo de tempo é independente do número de chegadas de qualquer outro intervalo de tempo disjunto. Por outro lado, se puder ser demonstrado analiticamente que os clientes chegam de forma independente uns dos outros, o processo de chegadas é um

processo de Poisson. Por este motivo, o processo de Poisson é considerado aleatório e é comum assumir que o número de chegadas por unidade de tempo pode ser estimado pela distribuição de probabilidade de Poisson, cuja expressão se apresenta a seguir [8].

$$P(n) = \frac{e^{-\lambda}.\lambda^{n}}{n!}, n = 0,1,2, \ldots \text{ onde} \begin{cases} P(n) \text{ é a probabilidade de } n \text{ chegadas} \\ n \text{ é o número de chegadas} \\ \lambda \text{ é a média de chegadas} \end{cases}$$

A distribuição exponencial e a distribuição de Poisson estão relacionadas, referindo-se a primeira ao tempo entre dois acontecimentos consecutivos e a segunda ao número de acontecimentos por unidade de tempo.

A função densidade de probabilidade (f.d.p) de uma variável aleatória contínua T, que pode representar o tempo entre chegadas ou o tempo de serviço, e que segue uma lei exponencial de parâmetro λ, é dada por

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, se \ \ t \geq 0 \\ 0 \quad\quad , se\ t < 0 \end{cases},$$

estando um exemplo do seu gráfico representado na Figura 2.5.

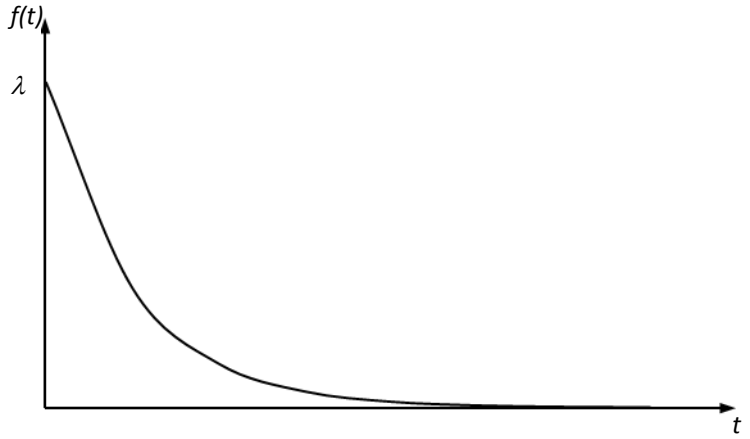


Figura 2.5 - Função densidade de probabilidade da lei exponencial

Resumindo, ao se considerar que o processo de Poisson é um modelo razoável para as chegadas num sistema de serviço específico, será importante considerar as suas três propriedades que o definem [8]:

- A probabilidade de chegar mais que um cliente num intervalo muito curto é quase nula;

- O número de clientes que chegam num intervalo de tempo é independente do número de clientes que chega em qualquer outro intervalo de tempo disjunto do primeiro;
- A probabilidade de um cliente chegar num intervalo muito curto é proporcional à amplitude desse intervalo.

Uma outra característica do padrão de chegada é a maneira como o padrão muda com o tempo, ou seja, se o padrão de chegada sofre ou não alterações ao longo do tempo. Se as chegadas forem independentes do tempo, quer dizer que a distribuição de probabilidade que descreve o comportamento de chegada no sistema é independente do tempo, designando-se por um padrão de chegada estacionário (*steady state*). Caso exista dependência do tempo, chama-se não estacionário (*transient*) [7].

A reação dos clientes ao entrarem no sistema também é importante, de maneira que, um cliente pode decidir esperar sem problema, independentemente do tamanho da fila, ou, por outro lado, o cliente pode decidir não entrar no sistema, caso a fila esteja muito grande. Assim, um cliente conhecido como impaciente é aquele que desiste de esperar ou simplesmente decide não entrar na fila se esta for muito grande. Um cliente paciente é aquele que permanece na fila até ser atendido. Nos sistemas onde existem duas ou mais filas, os clientes poderão ter um comportamento alternativo, optando por mudar de fila na esperança de ser atendido mais rapidamente noutro servidor.

No caso da urgência hospitalar, os utentes chegam independentemente uns dos outros, individualmente e o ritmo das chegadas não é constante, mas sim variável, aleatório, sendo, por isso, um processo estocástico. A observação e registo das chegadas permite avaliar qual a distribuição estatística que esta variável segue. Este registo pode ser feito através do número de utentes que chegam por unidade de tempo ou através do tempo entre chegadas sucessivas [5]. A reação dos utentes perante a fila de espera é, normalmente, paciente, ou seja, juntam-se à fila e esperam até serem atendidos.

2.2.1.3 Fila de espera

A fila de espera é constituída pelos clientes à espera de serem atendidos (não inclui o(s) cliente(s) em atendimento). Ela pode ser caracterizada quanto ao comprimento, número e disciplina da fila.

O comprimento da fila está relacionado com a capacidade do sistema e pode ter uma dimensão finita, por exemplo, devido a limitação física do espaço, podendo acolher apenas um número limitado de clientes ou pode ter dimensão infinita, quando não existe uma limitação ao tamanho da fila de espera.

Quanto ao número de filas de espera, pode existir uma fila única (fila simples) para um único servidor ou mesmo que o servidor tenha vários postos de atendimento (múltiplos servidores), ou podem existir múltiplas filas, sendo uma fila por posto de atendimento, onde cada conjunto fila/posto de atendimento constitui um sistema separado de fila de espera. Neste caso, é usual repartir as chegadas igualmente pelas várias filas. Estes casos estão representados na Figura 2.6.

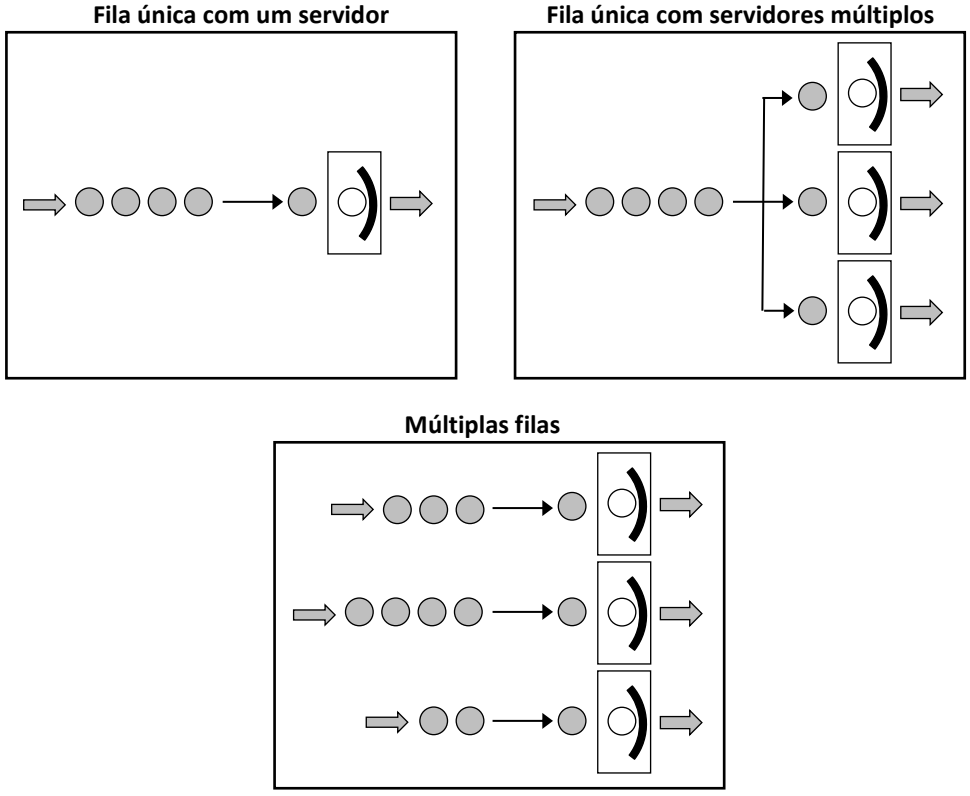


Figura 2.6 - Representação de Fila Única e de Múltiplas Filas

A disciplina da fila refere-se ao critério utilizado para definir a ordem de atendimento dos clientes que chegam ao sistema. A disciplina mais comum que pode ser observada na vida diária é FCFS (*First Come First Served*) ou FIFO (*First In First Out*), ou seja, o primeiro a chegar é o primeiro a ser servido, respeitando a ordem de chegada. Contudo, existem outras disciplinas, tais como, LCFS (*Last Come First Served*) ou LIFO (*Last In First Out*), aplicável, por exemplo, em sistemas de controlo de *stocks*, onde o item mais recente é mais fácil de ser servido, e ainda outras disciplinas baseadas em esquemas de prioridade, tais

como situações de reservas, idade ou emergências. A disciplina da fila pode ainda ser aleatória onde não existe nenhuma regra para que o cliente seja atendido.

No caso de um serviço de urgência hospitalar, não existe limite para a dimensão da fila, ou seja, existe a possibilidade de fila infinita. A disciplina da fila é, normalmente, FCFS, podendo, no entanto, ser estabelecida uma ordem de prioridade baseada no estado de saúde do utente, que indique que os utentes com determinadas características podem ir diretamente para o início da fila, aplicando, por exemplo, o sistema de triagem de Manchester, que se explica, resumidamente, no tópico seguinte.

2.2.1.3.1 Triagem de Manchester

A triagem é um sistema de prioridades, existente nas urgências hospitalares, que tem como objetivo o atendimento rápido de situações de risco para a saúde, sendo que quanto mais grave for a situação clínica do utente mais rapidamente ele deve ser atendido.

O sistema de triagem de Manchester utiliza um protocolo clínico que permite classificar a gravidade da situação de cada utente que recorre ao serviço de urgência. Após o utente efetuar a sua inscrição será encaminhado para um gabinete, onde será atendido por um profissional de saúde que lhe fará algumas perguntas sobre o motivo da sua vinda. Após uma observação rápida, objetiva e sistematizada é atribuída, ao utente, uma pulseira com uma determinada cor, em função da gravidade da doença.

Existem 5 categorias, representadas por 5 cores, vermelho, laranja, amarelo, verde e azul, cada uma representando um grau de gravidade e o tempo ideal em que o doente deverá ser atendido, tal como se mostra na Tabela 2.1.

Escala de Triagem de Manchester			
Nº	**Cor**	**Situação**	**Tempo alvo (minutos)**
1	Vermelho	Emergente	0
2	Laranja	Muito urgente	10
3	Amarelo	Urgente	60
4	Verde	Pouco urgente	120
5	Azul	Não urgente	240

Tabela 2.1 - Triagem de Manchester[1]

[1] Fonte: Grupo Português de Triagem (http://www.grupoportuguestriagem.pt).

Se o utente for considerado emergente (vermelho) entrará de imediato no balcão a que se destina. Se for considerado muito urgente (laranja) ou urgente (amarelo) entrará para uma sala de espera interna onde o médico o chamará para ser observado e tratado. Se for considerado pouco urgente (verde) ou não urgente (azul) aguardará na sala de espera a sua vez, que será quando não houver doentes mais graves para serem tratados.
A triagem de Manchester é um sistema baseado em fluxogramas. A primeira etapa diz respeito ao profissional de saúde que identifica a queixa principal do utente e a partir de 52 fluxogramas escolhe o mais adequado. Em seguida é desenvolvida uma entrevista estruturada e assinalada uma categoria que irá do nível 1 ao 5.

Refira-se também que existem diversos tipos de triagem. Escalas de triagem como a CTAS (*Canadian Triage and Acuity Scale*), ATS (*Australasian Triage Scale*), ESI (*Emergency Severity Index*), METTS (*Medical Emergency Triage and Treatment System*) e a MTS (*Manchester Triage Scale*) são amplamente divulgadas e instaladas nos diversos departamentos de emergência/urgência de todo o mundo. Em Portugal o sistema utilizado é a escala de triagem de Manchester.

2.2.1.4 Servidor

Relativamente à entidade que presta o serviço, o servidor, interessa conhecer o tempo que demora a efetuar o mesmo. O tempo de serviço é definido como sendo o tempo decorrido desde que o serviço foi solicitado até que este seja fornecido [9]. Como no caso das chegadas, o padrão de serviço pode ser prestado individualmente ou em grupo [7].

A capacidade dos sistemas de filas de espera é determinada pela capacidade de cada servidor e o número de servidores que estão a ser utilizados em paralelo (postos de atendimento), sendo, por isso, importante determinar o número de servidores e a sequência pela qual são visitados por cada cliente [17]. Quando se determina o número de servidores de um sistema, considera-se o número de servidores paralelos que poderão atender os clientes simultaneamente numa determinada fase do sistema [7].

A distribuição do tempo de serviço pode ser constante (por exemplo, se for um serviço prestado por uma máquina) ou aleatória. Por exemplo, nos serviços hospitalares, o tempo de serviço não é constante, mas antes variável, consoante o utente e quem presta o serviço. Geralmente, também se assume que cada servidor lida com um utente de cada vez [17].

Quanto maior for o grau de contacto do cliente com o prestador de serviço (servidor), mais difícil será o prestador de serviço controlar a duração do atendimento. No setor da saúde, em que a grande parte dos serviços prestados necessitam da presença do utente no sistema para que o serviço esteja concluído, a variabilidade dos tempos de atendimento é muito acentuada, variando de utente para utente, de médico para médico, de enfermeiro para enfermeiro.

No entanto, para modelar um sistema de filas de espera, por exemplo num serviço de urgências de um hospital, é necessário definir o tempo despendido pelo utente no servidor. Considerando que este tempo de serviço é uma variável aleatória, é necessário investigar sobre qual a distribuição estatística que lhe está associada. O caso das urgências tem, usualmente, como pressuposto que o tempo de serviço segue uma distribuição exponencial, com um determinado valor médio [5].

A capacidade média do servidor, medida em número de utentes servidos por unidade de tempo, obtém-se a partir do tempo médio de serviço, definindo-se assim a taxa de serviço habitualmente denotada por μ. Se, por exemplo, a duração média de uma consulta com um determinado médico for de 20 minutos, significa que a capacidade desse servidor (médico) é de 3 utentes por hora, ou seja, $\mu = 3$ utentes/hora.

A taxa de serviço pode ser independente do número de clientes existente no sistema ou dependente do número de clientes, ou seja, um servidor pode tornar-se mais ou menos eficiente consoante o tamanho da fila [7] e, nesse caso, se o número de clientes for, por exemplo n, escreve-se μ_n.

Um sistema de filas de espera pode ser formado por uma ou várias fases. Quando o cliente, ao longo da sua presença no sistema, tem de passar por mais do que um servidor diferente (que desempenha funções diferentes), então esse sistema de filas de espera tem múltiplas fases. Cada fase pode ser constituída por um ou vários servidores, formando-se fila única ou múltiplas filas.

Por exemplo, o serviço de urgências de uma unidade hospitalar terá um sistema com múltiplas fases, que serão a inscrição inicial, a triagem, a consulta, a realização de exames complementares de diagnóstico, etc. As figuras seguintes representam os diversos sistemas.

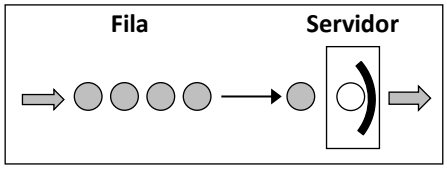


Figura 2.7 - Servidor único, fase única

Exemplo: Marcação de consulta

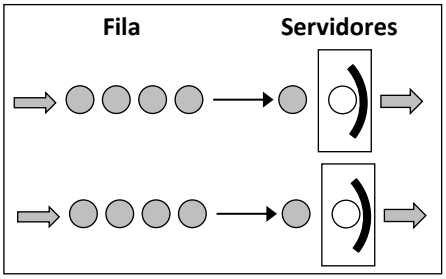
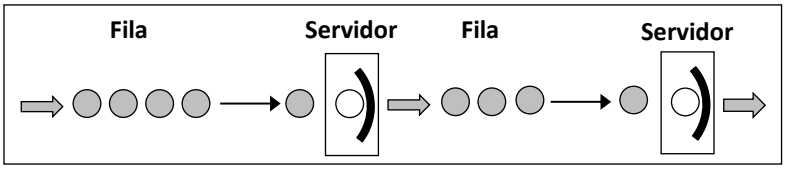
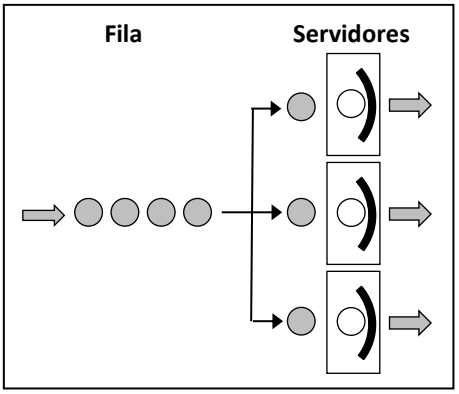


Figura 2.8 - Múltiplos servidores, uma fila por servidor, fase única



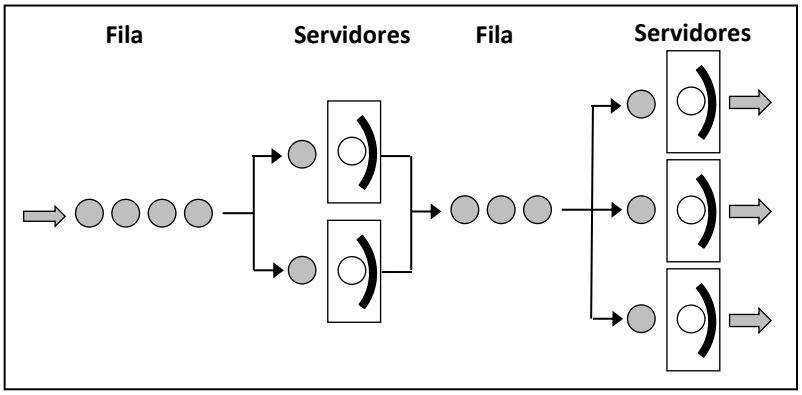
Exemplo: Consulta médica

Figura 2.9 - Servidor único, múltiplas fases (um servidor em cada fase)



Exemplo: Sala de tratamentos de enfermagem

Figura 2.10 - Múltiplos servidores, fase única



Exemplo: Serviço de urgência hospitalar com triagem

Figura 2.11 - Múltiplos servidores, múltiplas fases

Em modelos de filas de espera, o padrão de chegada e o tempo de serviço são modelados seguindo um processo estocástico baseado em distribuições de probabilidade [20]. Para formular um modelo de fila de espera como uma aproximação da realidade, a distribuição deve transcrever o comportamento real do sistema de forma mais fidedigna possível, mas deve permitir simultaneamente que o modelo produzido seja simples o suficiente para que possa ser tratado matematicamente [9].

2.3. Medidas de desempenho de um sistema de filas de espera

A gestão de um sistema de filas de espera, na maior parte dos casos, é uma tarefa complexa dado o carácter aleatório tanto das chegadas como do tempo de serviço. No caso concreto de uma unidade de prestação de cuidados de saúde, esta aleatoriedade pode ser reduzida recorrendo, por exemplo, a processos de marcação de consultas. Contudo esta medida não elimina a possibilidade de formação de fila de espera, pois não é possível reduzir a aleatoriedade no tempo do serviço.

A gestão de filas de espera passa por tomar medidas em relação aos três componentes principais de um sistema: chegadas, fila de espera e servidor. Por isso, deverá configurar-se o sistema de filas de espera de forma a encontrar um equilíbrio entre a qualidade do serviço prestado e os custos inerentes a esse serviço. Assim, a melhor configuração deve passar pela tomada de decisão em relação ao número de servidores a instalar, número de fases, tipo de fila (fila única para vários servidores ou uma fila por servidor), segmentação ou não dos clientes, usando, por exemplo, servidores segmentados por tipo de cliente. Todas estas decisões irão ter impacto na qualidade do serviço prestado, em função do tempo de espera, e no seu custo associado [5].

É necessário encontrar soluções equilibradas, o que implica estabelecer *trade-offs* entre custo e qualidade de serviço. Uma solução de custo mínimo, por exemplo, apenas com um servidor, poderá comprometer a qualidade do serviço, originando um tempo de espera elevado. Uma solução de qualidade elevada, de modo a originar um tempo de espera reduzido, implicará custos elevados, por exemplo, devido à existência de vários servidores.

A Figura 2.12, baseada em [5], representa graficamente as funções do tempo de espera e do custo do serviço fundamentadas pela explicação anterior.

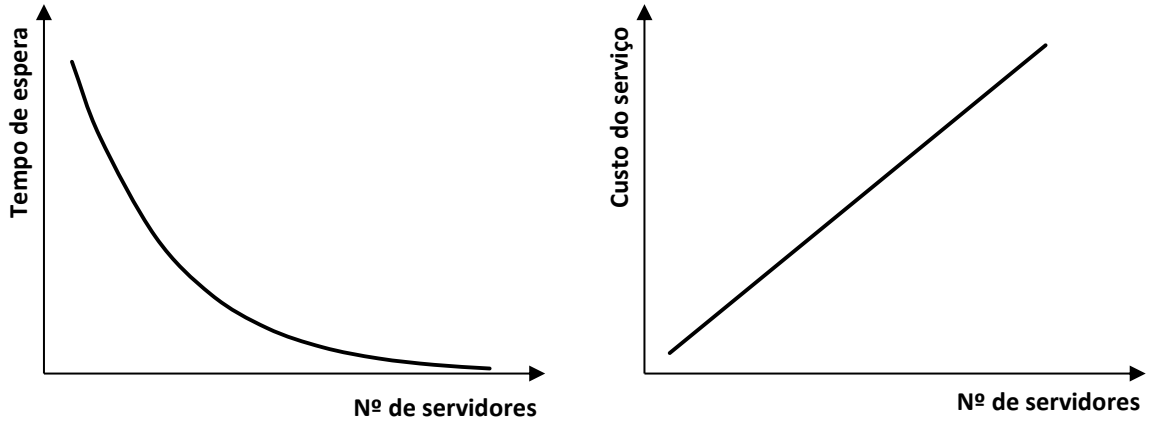


Figura 2.12 - Funções do tempo de espera e do custo do serviço consoante o n.º de servidores

Considera-se que, no início do seu funcionamento, o sistema atravessa um estado transitório que poderá evoluir para um estado estacionário. Assume-se, ao longo deste estudo, que se verificam as condições necessárias para que o sistema atinja o estado estacionário. Neste estado, as medidas de desempenho não dependem de há quanto tempo o sistema está em funcionamento nem do estado inicial do sistema, verificando-se que o número médio de entradas é igual ao número médio de saídas e a distribuição de probabilidade mantém-se a mesma ao longo do tempo.

A configuração de um sistema de filas de espera deve ser avaliada através de algumas medidas de desempenho, capazes de dar uma indicação sobre a qualidade do serviço esperada e o respetivo custo. O tópico seguinte dará relevo à notação utilizada nas medidas necessárias para avaliar o desempenho de sistema de filas de espera.

2.3.1 Terminologia aplicada às medidas de desempenho

A literatura sobre a teoria das filas de espera é quase unânime na adoção de uma terminologia para a quantificação de sistemas de filas de espera. As tabelas seguintes indicam os símbolos usados para a especificação do modelo e para as medidas de desempenho.

Especificação do modelo	
Símbolo	**Característica**
λ	**Taxa média de chegada dos clientes**
$1/\lambda$	Tempo médio entre chegadas
μ	**Taxa média de serviço de um servidor**
$1/\mu$	Tempo médio de serviço a um cliente
s	**Número de servidores**
ρ	**Taxa de utilização do sistema**
$\rho = \lambda/\mu$	Taxa de utilização do servidor
$1 - \rho$	Taxa média de desocupação do serviço

Tabela 2.2 - Terminologia usada na especificação do modelo

Ainda relativamente à especificação do modelo será importante caracterizar a estrutura da fila de espera (fila única ou múltiplas filas). Será este elemento que vai ditar qual dos modelos deve ser aplicado, se o modelo de servidor simples ou o modelo de múltiplos servidores.

Será necessário conhecer as características referidas na Tabela 2.2 para o cálculo das medidas de desempenho no regime estacionário, mencionadas na Tabela 2.3 [9]. O sistema está em regime estacionário quando a distribuição de probabilidade do sistema se mantém a mesma ao longo do tempo.

Medidas de desempenho	
Símbolo	**Característica**
L_s	Número médio de clientes no sistema
L_q	Número médio de clientes na fila de espera
W_s	Tempo médio que um cliente está no sistema (tempo na fila + tempo de serviço)
W_q	Tempo médio que um cliente está na fila de espera (exclui o tempo que o cliente demora a ser atendido)
P_n	Probabilidade de estarem n clientes no sistema
P_0	Probabilidade de estarem zero clientes no sistema
$P(W_q = 0)$	Probabilidade do tempo de espera na fila ser zero
$P(W_q > t)$	Probabilidade do tempo de espera na fila exceder t
$P(W_s > t)$	Probabilidade do tempo no sistema exceder t

Tabela 2.3 - Terminologia usada nas medidas de desempenho

A letra *L* é usada em questões relacionadas com o comprimento da fila (L da palavra inglesa *Length* - Comprimento) e a letra *W* é usada em questões relacionadas com tempo

de espera (W da palavra inglesa *Waiting* - esperando). Da mesma forma, usam-se as letras *q* e *s* para indicar a fila (*queue* em inglês) e o sistema (*system* em inglês), respetivamente.

A taxa média de utilização do servidor pode ser utilizada como um indicador do custo de serviço e o tempo de espera como um indicador da qualidade do serviço. A relação entre estes dois indicadores encontra-se representada na Figura 2.13 [5].

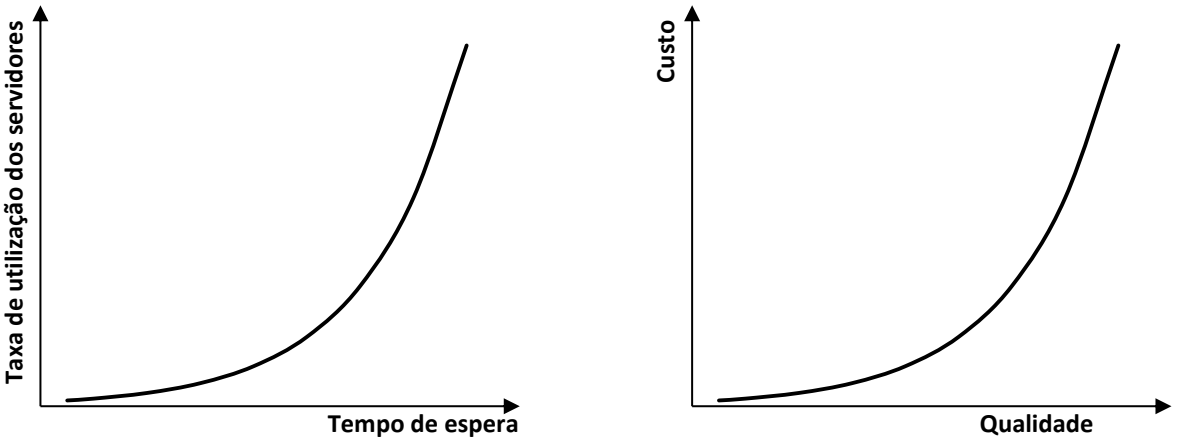


Figura 2.13 - Relação da taxa média de utilização com o tempo de espera

Num sistema de filas de espera, seja 1/ λ o tempo médio entre duas chegadas consecutivas, L_s o número médio de unidades no sistema, e W_s o tempo médio que cada unidade despende no sistema. A Lei de Little diz que, se as três médias são finitas e o processo estocástico correspondente é estacionário, então $L = \lambda . W$, para a fila e para o sistema [14].

Existem cinco relações chave que fornecem a base para a formulação de filas de espera com modelos de fonte infinita [17]:

- Taxa de utilização do sistema [2]: $\rho = \frac{\lambda}{s\mu}$
- Número médio de clientes no sistema: $L_s = L_q + \rho$
- Tempo médio na fila: $W_q = \frac{L_q}{\lambda}$
- Tempo médio no sistema: $W_s = W_q + \frac{1}{\mu}$

[2] $s\mu$ é a capacidade do serviço, sendo s o número de servidores. Se $s=1$, então $\rho=\lambda/\mu$.

Na Figura 2.14, apresenta-se um resumo da terminologia onde *n* representa o número de clientes no sistema e *N(t)* o número de clientes no instante *t* ($t \geq 0$). As outras variáveis já foram mencionadas anteriormente.

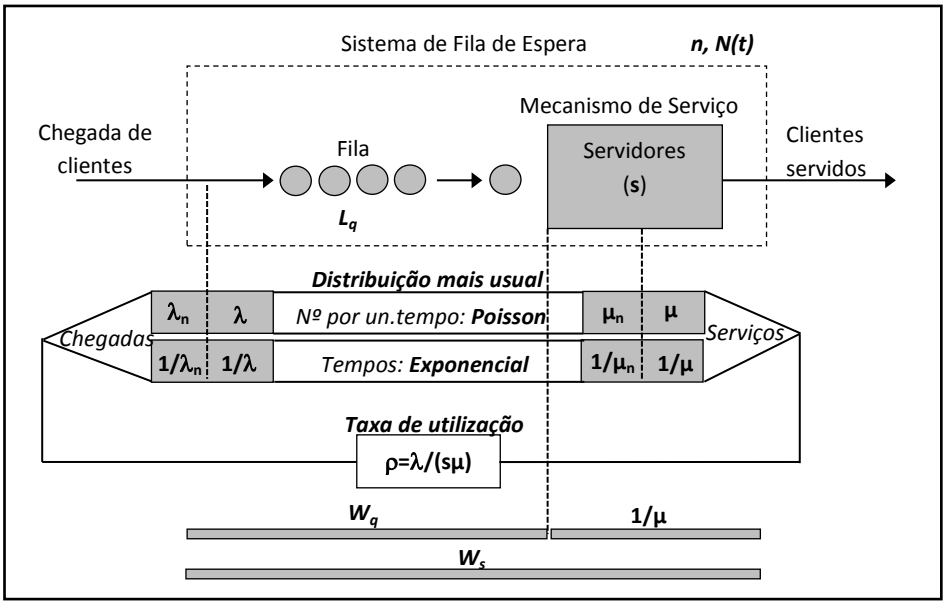


Figura 2.14 - Resumo da terminologia

2.4. Modelos das filas de espera

A maioria dos modelos analíticos de filas de espera supõe chegadas seguindo uma distribuição de Poisson e um tempo de atendimento caracterizado por uma distribuição exponencial. Estas são as distribuições que mais frequentemente caracterizam as filas de espera reais e têm duas propriedades importantes:

- Se o tempo entre dois acontecimentos consecutivos segue uma distribuição exponencial com parâmetro λ, então o número de acontecimentos por unidade de tempo t segue uma distribuição de Poisson com parâmetro λt.
- A distribuição exponencial não tem memória, isto é, a probabilidade de ocorrência de um acontecimento é independente do instante de tempo em que ocorreu o acontecimento imediatamente anterior.

A notação padrão usada nos modelos para descrever muitos dos sistemas de filas de espera foi desenvolvida por Kendall, em 1951, onde cada sistema de filas é descrito por seis características [2].

2.4.1 Notação de Kendall

As características abordadas anteriormente descrevem um sistema de filas de espera. Para simplificar os diversos modelos das filas utiliza-se a notação de Kendall composta por seis características da seguinte forma [15]: **A/S/m/K/N/Q**

Estas características estão explicadas na Tabela 2.4.

Notação de Kendall [24]	
Símbolo	**Característica**
A	**Distribuição dos tempos entre as chegadas:** representa a distribuição do intervalo de tempo entre chegadas consecutivas
M	Tempos entre chegadas são independentes e identicamente distribuídos (iid), variáveis aleatórias com uma distribuição exponencial
D	Tempos entre chegadas são iid e determinísticos
E_k	Tempos entre chegadas são iid seguindo uma lei de Erlang com parâmetro k
G	Tempos entre chegadas são iid e são regulados por uma distribuição geral
S	**Distribuição dos tempos de serviço:** caracteriza a duração do tempo de serviço
M	Tempos de serviço são iid com distribuição exponencial
D	Tempos de serviço são iid e determinísticos
E_k	Tempos de serviço são iid seguindo uma lei de Erlang com parâmetro k
G	Tempos de serviço são iid e seguem uma distribuição geral
m	**Número de servidores em paralelo**
K	**Capacidade do sistema:** define o n.º máximo de clientes permitidos no sistema que, por norma, é infinito e representa os clientes que estão na fila e os que estão a ser servidos
N	**Tamanho da população:** normalmente é considerado infinito, a não ser que o número de potenciais clientes seja da mesma ordem de grandeza que o número de servidores
Q	**Disciplina da fila**
FCFS	*First Come, First Served* (primeiro a entrar, primeiro a ser servido)
LCFS	*Last Come, First Served* (último a entrar, primeiro a ser servido)
SIRO	*Service In Random Order (serviço em ordem aleatória)*
NPRP	*Nonpreemptive Priority Models* (prioridade “não absoluta”)
PRP	*Preemptive Priority Models* (prioridade “absoluta”)
GD	Disciplina geral da fila

Tabela 2.4 - Notação de Kendall

Em muitos modelos importantes, as características K/N/Q são do tipo ∞/∞/FCFS. Quando for este o caso, estas características podem ser omitidas. Esta regra será aplicada na maior parte das explicações seguintes dos diversos tipos de modelos, ou seja, será assumido que, nos diversos casos, o sistema tem capacidade infinita, o tamanho da população é também infinito e a fila possui uma disciplina de atendimento do tipo FCFS.

Na Tabela 2.5, estão representados alguns exemplos de modelos, seguindo a notação referenciada anteriormente.

M/D/1:	Tempos entre chegadas com distribuição exponencial; tempos de serviço constantes (determinísticos); servidor único.
E_k/G/2:	Tempos entre chegadas seguem uma lei de Erlang; tempos de serviço com uma distribuição geral; 2 servidores.
G/E_k/s:	Tempos entre chegadas com distribuição geral; tempos de serviço seguem uma lei de Erlang; *s* servidores.
M/M/3/9/LCFS:	Tempos entre chegadas e tempos de serviço com distribuição exponencial; 3 servidores; limite de 9 clientes dentro do local de prestação de serviços ao mesmo tempo, sendo que o último cliente a chegar é o primeiro a ser atendido.

Tabela 2.5 - Exemplos de modelos seguindo a notação de Kendall

2.4.2 Processo de nascimento e morte

A maior parte dos modelos elementares de filas de espera baseiam-se no processo de nascimento e morte. No contexto das filas de espera, um nascimento corresponde à chegada de um novo cliente e uma morte corresponde à saída de um cliente do sistema com o serviço concluído [9].

O estado do sistema no instante t é o número de clientes no sistema, denotado por *N(t)*. Um processo de nascimento e morte obedece a três hipóteses base:

Hipótese 1: Dado $N(t) = n$, a distribuição de probabilidade do tempo restante até ao próximo nascimento (chegada) é exponencial com parâmetro λ_n (n = 0, 1, 2, …).

Hipótese 2: Dado *N(t) = n*, a distribuição de probabilidade do tempo restante até à próxima morte (final de atendimento) é exponencial com parâmetro μ_n (n = 0, 1, 2, …).

Hipótese 3: Nunca podem ocorrer, em cada instante, mais do que um nascimento ou uma morte (transição apenas para estados adjacentes).

As hipóteses 1 e 2 tornam o processo de nascimento e morte um tipo particular da Cadeia de Markov (os estados futuros do sistema são independentes do passado e dependem exclusivamente do estado atual), o que facilita o tratamento das filas de espera que assim podem ser descritas. A hipótese 3 simplifica adicionalmente a análise.

Na Figura 2.15, esquematiza-se o diagrama de transição correspondente ao processo de nascimento e morte, onde cada estado é identificado pelo número de clientes no sistema.

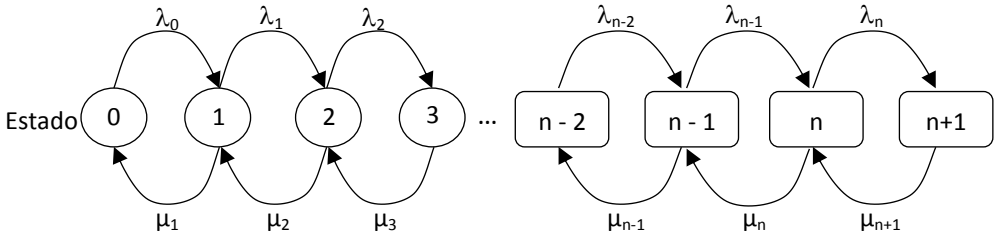


Figura 2.15 - Diagrama de transição correspondente ao processo de nascimento e morte

λ_n: taxa média de chegadas (número esperado de chegadas por unidade de tempo) de novos clientes quando *n* clientes estão no sistema.

μ_n: taxa média de serviço (número esperado de atendimentos concluídos por unidade de tempo) quando *n* clientes estão no sistema.

As setas no diagrama representam as únicas transições possíveis no estado do sistema e os valores inscritos por cima, ou por baixo, de cada seta representam a respetiva taxa média para essa transição, assumindo que o sistema atingiu um estado estacionário.

Os processos de nascimento e morte são processos de Markov de espaço discreto no qual as transições entre estados estão restritas aos estados vizinhos.

Nos tópicos seguintes, apresentam-se alguns dos modelos mais aplicados no estudo de sistemas de filas de espera, baseados no processo de nascimento e morte, considerando, nos casos mais comuns, os pressupostos apresentados na Tabela 2.6.

- Distribuição de Poisson para as chegadas, com λ = taxa média de chegadas;
- Distribuição Exponencial para o tempo de serviço, com μ = taxa média de serviço;
- Disciplina da fila: FCFS (primeiro a chegar é o primeiro a ser atendido);
- Todas as chegadas esperam em fila até serem servidas;
- Possibilidade de extensão infinita da fila.

Tabela 2.6 - Pressupostos para aplicação dos modelos mais comuns

2.4.3 Modelo M/M/s

Neste modelo, assume-se que os intervalos de tempos entre chegadas consecutivas são independentes e identicamente distribuídos, com distribuição exponencial (**M**), que os tempos de serviço são independentes e identicamente distribuídos, com distribuição exponencial (**M**) e que existem múltiplos servidores (**s**), que fornecem serviço independente uns dos outros. Como as últimas três características da referência do modelo estão omissas, assume-se que o sistema tem capacidade ilimitada, que é alimentado por uma população infinita e que a disciplina da fila praticada será FCFS.

A Figura 2.16 representa esquematicamente, de uma forma resumida, este modelo.

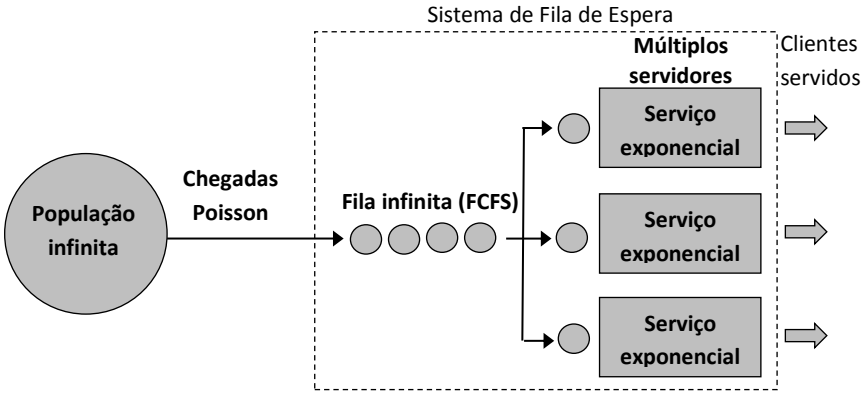


Figura 2.16 - Modelo M/M/s

Se existir uma fila para vários servidores, então o modelo a aplicar será o modelo de múltiplos servidores. Será, por exemplo, o caso de um serviço de urgências de uma

unidade de cuidados de saúde, onde os utentes fazem a sua inscrição quando entram nas urgências, formando uma única fila para depois serem distribuídos pelos vários servidores existentes.

De notar que, se a taxa média de chegadas de clientes ao sistema continua a ser λ, independentemente do estado, então a taxa média de serviço, que se assume ser μ por cada um dos *s* servidores, dependerá do estado do sistema. Sendo assim, quando a taxa de serviço média por servidor ocupado é μ, a taxa de serviço média geral para *n* servidores ocupados deve ser $n.\mu$, ou seja, quanto mais servidores estiverem ocupados mais provável é sair um cliente atendido do sistema. Portanto, $\mu_n = n.\mu$, quando $n < s$, enquanto que $\mu_n = s.\mu$ quando $n \geq s$ (todos os s servidores ocupados) [9].

$$\mu_n = \begin{cases} n.\mu \,;\, se\ n < s \\ s.\mu \,;\, se\ n \geq s \end{cases}$$

As medidas de desempenho dos sistemas de filas de espera com a configuração atrás mencionada são calculadas com base nas fórmulas referidas na Tabela 2.7 [5].

Modelo M/M/s	
Significado	**Fórmula**
Taxa de utilização do sistema (ρ)	$\rho = \frac{\lambda}{s.\mu};\ \rho < 1$
N.º médio de clientes na fila (L_q)	$L_q = \frac{P_0 \left(\frac{\lambda}{\mu}\right)^s \rho}{s!\,(1-\rho)^2}$
N.º médio de clientes no sistema (L_s)	$L_s = L_q + \frac{\lambda}{\mu}$
Tempo médio na fila de espera (W_q)	$W_q = \frac{L_q}{\lambda}$
Tempo médio no sistema (W_s)	$W_s = W_q + \frac{1}{\mu} = \frac{L_s}{\lambda}$
Probabilidade de 0 clientes no sistema ou Probabilidade do tempo de espera ser 0 ou Taxa de inatividade do servidor (P_0)	$P_0 = \frac{1}{\left[\sum_{n=0}^{s-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}\right] + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} \times \left(1 - \frac{\lambda}{s\mu}\right)^{-1}}$
Probabilidade de *n* clientes no sistema se 1 ≤ n ≤ s (P_n)	$P_n = P_0 \left[\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}\right]$
Probabilidade de *n* clientes no sistema se n ≥ s (P_n)	$P_n = P_0 \left[\frac{\left(\frac{\lambda}{\mu}\right)^n}{s!\,s^{n-s}}\right]$
Probabilidade do tempo de espera na fila exceder t ($P_{(Wq > t)}$)	$P_{W_q > t} = (1 - P_0) \times e^{-s\mu(1-\rho)t}$
Probabilidade do tempo no sistema exceder t ($P_{(Ws > t)}$)	$P_{W_s > t} = e^{-\mu t}\left[\frac{1 + P_0\left(\frac{\lambda}{\mu}\right)^s}{s!\,(1-\rho)}\left(\frac{1 - e^{-\mu t\left(s-1-\lambda/\mu\right)}}{s-1-\lambda/\mu}\right)\right]$

Tabela 2.7 - Fórmulas do modelo M/M/s

A condição para que a fila de espera atinja o estado estacionário de modo que não cresça sem limite é $s.\mu > \lambda$ ou, equivalentemente, $\rho < 1$, uma vez que $\rho = \lambda/(s.\mu)$. Por outras palavras, poder-se-á dizer que a taxa média de chegada deve ser menor que a taxa média máxima de serviço do sistema, que é intuitivamente o que seria esperado.

2.4.4 Modelo M/M/1

Neste modelo, assume-se que os intervalos de tempos entre chegadas consecutivas são independentes e identicamente distribuídos, com distribuição exponencial (**M**), que os tempos de serviço são independentes e identicamente distribuídos, com distribuição exponencial (**M**) e que existe um único servidor (**1**). Como as últimas três características estão omissas, assume-se que o sistema tem capacidade ilimitada, que é alimentado por uma população infinita e que a disciplina da fila praticada será FCFS.

A Figura 2.17 representa esquematicamente, de uma forma resumida, este modelo.

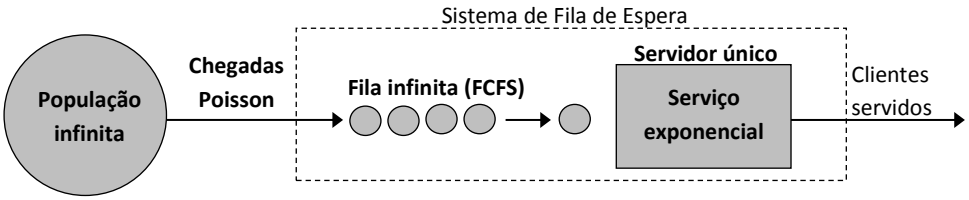


Figura 2.17 - Modelo M/M/1

Se existir uma fila por servidor, então o modelo a aplicar será o modelo de servidor simples. Esta situação pode ocorrer quando existe apenas um servidor ou quando existem múltiplas filas independentes umas das outras, havendo um servidor para cada fila (ver Figura 2.8).

Quando o sistema tem mais do que uma fase, cada fase deve ser analisada em separado, identificando qual o modelo a aplicar em cada uma. Será, por exemplo, o caso de um serviço de urgências de uma unidade de cuidados de saúde, onde os utentes, numa primeira fase, fazem a sua inscrição quando entram nas urgências e depois esperam até serem chamados para a triagem, que corresponderá à segunda fase.

Por ser um caso particular do modelo anterior, as medidas de desempenho, dos sistemas de filas de espera com a configuração M/M/1, são calculadas com base nas fórmulas do modelo M/M/s, fazendo $s = 1$.

O uso dos resultados pressupõe que $\mu > \lambda$, isto é, $\rho < 1$, caso contrário a fila de espera cresceria sem limite e o sistema nunca atingiria o estado estacionário. Dito de outro modo,

para que o sistema atinja o estado estacionário, a taxa de chegadas terá que ser menor que a taxa de serviço para que a fila não cresça infinitamente [13].

Neste modelo, como existe apenas um servidor, o sistema pode tolerar a utilização a 100%. Se a taxa de chegada for maior do que a taxa de serviço, então, um modelo com múltiplos servidores será o mais apropriado.

2.4.5 Modelo M/M/s/K

Um sistema M/M/s/K é um sistema de capacidade finita com **s** servidores com tempos de atendimento exponenciais independentes e identicamente distribuídos (os quais não dependem do estado do sistema). Como a capacidade do sistema deve ser, no mínimo, tão grande quanto o número de servidores, s ≤ K. Para tal sistema [9],

$$\lambda_n = \begin{cases} \lambda\,; n = 0,1,\dots,K-1 \\ 0;\ \ n = K, K+1,\dots \end{cases}$$

À semelhança do que aconteceu no modelo M/M/s, a taxa média de saída de cada estado também será dependente do estado. Sendo assim,

$$\mu_n = \begin{cases} n.\mu\,; se\ n < s \\ s.\mu\,; se\ n \geq s \end{cases}$$

O modelo M/M/s/K tende para o modelo M/M/s quando K tende para infinito.

Algumas das medidas de desempenho dos sistemas de filas de espera com a configuração atrás mencionada são calculadas com base nas fórmulas referidas na Tabela 2.8 [3].

Modelo M/M/s/K *s ≤ K; Nº de servidores = s; Nº máximo de clientes no sistema = K; População = ∞; Fila FCFS*	
Significado	**Fórmula**
Taxa de utilização do sistema (ρ)	$\rho = \frac{\lambda}{s\mu}$
Taxa média de entrada	$\bar{\lambda} = \lambda(1 - P_K)$
Taxa de ocupação	$\frac{\bar{\lambda}}{s\mu}$
Taxa média de entrada	$\bar{\lambda} = \lambda(1 - P_K)$
N.º médio de clientes na fila (L_q)	$L_q = \frac{s^s \rho^{s+1}}{s!\,(1-\rho)^2}[1 - \rho^{K-s} - (1-\rho)(K-s)\rho^{K-s}]P_0$
N.º médio de clientes no sistema (L_s)	$L_s = L_q + \frac{\bar{\lambda}}{\mu}$
Tempo médio na fila de espera (W_q)	$W_q = \frac{L_q}{\bar{\lambda}}$
Tempo médio no sistema (W_s)	$W_s = W_q + \frac{1}{\mu} = \frac{L_s}{\bar{\lambda}}$
Probabilidade de 0 clientes no sistema ou Probabilidade do tempo de espera ser 0 ou Taxa de inatividade do servidor (P_0)	$P_0 = \left[\frac{s^s \rho^{s+1}(1-\rho^{K-s})}{s!\,(1-\rho)} + \sum_{n=0}^{s} \frac{(s\rho)^n}{n!}\right]^{-1}; \rho \neq 1$ $P_0 = \left[\frac{s^s}{s!}(K-s) + \sum_{n=0}^{s} \frac{s^n}{n!}\right]^{-1}; \rho = 1$
Probabilidade de *n* clientes no sistema (P_n)	$P_n = \frac{(s\rho)^n}{n!}P_0;\ n \leq s$ $P_n = \frac{s^n \rho^n}{s!}P_0;\ n = s+1, \ldots, K$ $P_n = 0;\ n > K$

Tabela 2.8 - Fórmulas do modelo M/M/s/K

2.4.6 Modelo M/M/1/K

Este modelo difere do modelo M/M/1 porque tem capacidade do sistema limitada até **K** clientes ao mesmo tempo no local da prestação do serviço. Quando já estiverem K clientes no sistema e, caso se verifique a chegada de um novo cliente, ser-lhe-á recusado o acesso ao sistema, ou seja, trata-se de uma fila de espera com capacidade finita. De notar que os potenciais clientes com acesso vedado não poderão aguardar no exterior do sistema, para

entrada posterior. Este modelo pode, por exemplo, traduzir situações em que a sala de espera tem uma dimensão limitada.

Se λ designar a taxa de chegadas de clientes ao local de prestação de serviços, então a taxa de entradas no sistema será dependente do estado *n*, isto é, depende do número de clientes no sistema e será dada por [3]:

$$\lambda_n = \begin{cases} \lambda\,; n = 0,1,\dots,K-1 \\ 0;\ \ n = K, K+1, \dots \end{cases}$$

A taxa média de entrada no sistema, normalmente denotada por $\bar{\lambda}$, é a média ponderada das taxas λ e é dada por:

$$\bar{\lambda} = \lambda \times \sum_{n=0}^{K-1} P_n = \lambda(1 - P_K).$$

Por ser um caso particular do modelo anterior, as medidas de desempenho dos sistemas de filas de espera com a configuração M/M/1/K são calculadas com base nas fórmulas do modelo M/M/s/K, fazendo *s = 1*.

Note-se que a probabilidade de o sistema estar num estado n ≥ K+1 é nula uma vez que não se aceita ninguém acima de K.

O sistema pode estar em equilíbrio para valores de ρ superiores a 1. No entanto, nesse caso, haverá um número potencialmente elevado de clientes do sistema que chegam e não são servidos.

2.4.7 Modelo M/M/∞

Este modelo corresponde a um modelo de filas no qual o serviço é ilimitado, isto é, existe um número infinito de servidores disponíveis, não existindo, por isso, fila de espera [1]. Para além disso, a distribuição dos tempos entre as chegadas e dos tempos de atendimento é exponencial, a capacidade do sistema e da população são infinitas e a disciplina da fila segue a ordem FCFS. Um exemplo simples é o caso de um hipermercado onde os clientes se servem a eles próprios.

Os clientes ao entrarem no sistema são servidos de imediato, sendo o tamanho esperado do sistema a média da distribuição de Poisson, obtido de $L_s=\rho=\lambda/\mu$. Como o número de servidores é infinito, $L_q = 0 = W_q$. O tempo médio no sistema será o tempo médio de serviço, de modo que $W_s = 1/\mu$ e a função de distribuição do tempo de espera é idêntica à distribuição do tempo de serviço, ou seja, exponencial com média $1/\mu$.

Por este facto, o tempo de espera no sistema é inversamente proporcional à taxa de serviço, ou seja, enquanto a taxa de serviço aumenta, diminui o tempo de espera no serviço, como seria expectável.

2.4.8 Modelos que envolvem distribuições não exponenciais

Embora a distribuição exponencial descreva, em muitas aplicações, o processo de chegada, há muitas situações em que talvez esta não se encaixe muito bem no processo do serviço. Contudo, existem generalizações do modelo básico que permitem que a distribuição do tempo de serviço seja arbitrária.

Todos os modelos da teoria de filas de espera abordados anteriormente basearam-se no pressuposto de que tanto os tempos entre chegadas como os tempos de serviços tinham distribuições exponenciais. Para determinados tipos de sistemas de filas, esta distribuição, apenas nos dará uma aproximação à análise do estudo que pretendemos implementar. Por exemplo, assumir que os tempos entre chegadas têm uma distribuição exponencial, requer que as chegadas se sucedam aleatoriamente (processo de Poisson). Esta situação pode ser aceitável na maioria dos casos, mas não será coerente quando as chegadas são programadas. Será, por isso, importante, quando os modelos não têm distribuições exponenciais, analisar outros modelos de filas alternativos [9].

Os modelos de filas de espera que não seguem as distribuições exponenciais têm uma análise matemática mais difícil apesar de já se ter conseguido alguns resultados aplicáveis a alguns destes modelos [9].

Nos tópicos seguintes será feito um resumo de alguns modelos com distribuição não exponencial.

2.4.8.1 Modelo M/G/1

Seguindo a notação de Kendall este modelo pressupõe que o sistema de filas tem um único servidor e um processo de chegadas com distribuição exponencial. A taxa média de chegada continua a ser dada por λ e, tal como nos modelos anteriores, os tempos de serviço são independentes.

O que difere neste modelo é o facto de não existir nenhuma restrição no que diz respeito à distribuição do tempo de serviço. A referência indica que é uma distribuição generalizada (**G**). Neste caso, apenas será necessário conhecer $1/\mu$ e a variância σ^2 dessa distribuição [9].

Se $\rho = \lambda / \mu < 1$ o sistema poderá eventualmente atingir o estado de equilíbrio, continuando a ser válidas as relações principais que fornecem a base para a formulação dos modelos de filas de espera. A partir da fórmula de L_q (número médio de clientes na fila), dada por

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)},$$

é possível obter todos os outros resultados analíticos.

Se a duração do atendimento for exponencial, $\sigma^2 = 1/\mu^2$, obtêm-se os resultados do modelo M/M/1.

2.4.8.2 Modelo M/D/1

Se continuarmos a assumir que o processo de chegadas segue uma distribuição de Poisson (**M**), que temos apenas um servidor e se o tempo de serviço aos diferentes clientes consistir numa rotina relativamente idêntica, sendo constante ou determinístico (**D**), praticamente não se verificará qualquer variação na duração do serviço, sendo então útil o modelo M/D/1, que pode ser encarado como um caso particular do modelo M/G/1, fazendo $\sigma^2 = 0$.

Deste modo, a fórmula para determinar o número médio de clientes na fila será dada por [9]

$$L_q = \frac{\rho^2}{2(1-\rho)}.$$

Os valores de L_s, W_q e W_s podem ser obtidos pelas expressões do modelo anterior.

É interessante notar que o valor indicado na fórmula acima é metade do correspondente valor para o modelo M/M/1, ou seja, se a distribuição do atendimento for exponencial com parâmetro μ, o comprimento da fila de espera será duplo do que seria se todos os atendimentos fossem executados com duração determinística (igual a $1/\mu$). Fica assim patente a importância da variância da distribuição da duração do atendimento no desempenho do sistema.

2.4.8.3 Modelo M/Ek/1

Como se referiu anteriormente, nos sistemas M/D/1 assume-se que a duração do atendimento de um cliente é determinística ($\sigma = 0$), uma situação teórica que raramente ocorre rigorosamente na prática. Num outro extremo, temos os modelos M/M/s, em que se assume uma variação muito grande ($\sigma = 1/\mu$ = duração média do atendimento de um cliente). Ora, na realidade, muitas vezes nem temos uma duração determinística, nem temos uma variação tão elevada. É para estes casos que se torna útil recorrermos à distribuição Erlang-k.

A distribuição de Erlang é uma distribuição de probabilidade contínua que é igual ao somatório de um número de *k* variáveis aleatórias exponenciais independentes. O parâmetro *k* representa o número de fases do sistema e quando este é composto por uma sequência de serviços, cada um deles com uma distribuição exponencial, o tempo total de serviço tem uma distribuição de Erlang.

Considere-se um sistema com um processo de Poisson nas chegadas, com taxa λ, e com a duração do atendimento de um cliente, T, com média $1/\mu$. Imagine-se que se sabe que a duração de cada atendimento não segue uma distribuição exponencial e que se assume que cada atendimento se pode decompor numa sequência de k estádios consecutivos, cada um deles com durações, T_i, independentes e identicamente distribuídas, com distribuição exponencial de valor médio $1/(k\mu)$.

A distribuição Erlang-k, com parâmetros *k* e μ, tem valor médio igual a $1/\mu$ e variância igual a $1/(k\mu^2)$. Assim, o coeficiente de variação da distribuição Erlang-k será [9]

$$CV_{E_k} = \frac{\frac{1}{\sqrt{k}.\mu}}{\frac{1}{\mu}} = \frac{1}{\sqrt{k}}.$$

De notar que o coeficiente de variação da distribuição Erlang-k é sempre menor ou igual a 1. A igualdade ocorre quando *k = 1*, ou seja, quando a distribuição Erlang-k coincide com a distribuição exponencial, como se pode ver pela Figura 2.18.

A função densidade de probabilidade de uma variável *t* com uma distribuição de Erlang, com parâmetros *k* e μ, é dada por [1, p. 17]

$$f(t) = \frac{(\mu k)^k . t^{k-1} . e^{-k\mu t}}{(k-1)!}, \quad para \quad t, \mu, k \geq 0 \wedge k \in \mathbb{N}.$$

Um exemplo gráfico da função, para vários valores de *k*, está representado na Figura 2.18, adaptada de [9, p. 875].

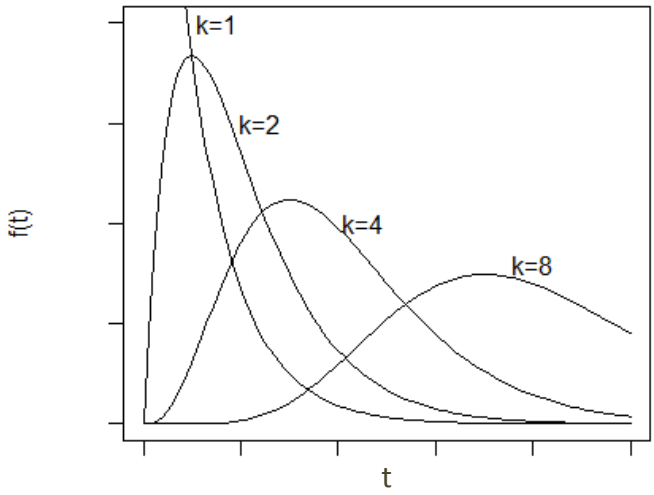


Figura 2.18 - Família de distribuições de Erlang com média constante de *1/μ*

O Modelo $M/E_k/1$ pode ser caracterizado a partir da análise do correspondente diagrama de transição de estados, baseado no processo de nascimento e morte.

Poderíamos escrever as equações de equilíbrio para os vários estados e, após várias manipulações, deduzir alguns resultados. No entanto, como já apresentámos o modelo mais geral M/G/1, poderemos encarar o modelo $M/E_k/1$ como um caso particular desse, com $\sigma^2 = 1/(k\,\mu^2)$.

Assim, a fórmula para a determinação do número médio de clientes na fila será [9]:

$$L_q = \frac{1+k}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)}$$

Os outros parâmetros relevantes podem ser obtidos aplicando as fórmulas apresentadas no modelo M/G/1.

2.4.9 Modelos sem entradas com distribuição de Poisson

Os modelos M/…/… assumem um processo de chegadas com distribuição de Poisson (intervalos de tempo entre chegadas consecutivas independentes e identicamente distribuídos, com distribuição exponencial). No entanto, em certas situações, tal poderá não ser o mais adequado, como é o caso de as chegadas estarem previamente agendadas.

Se a duração do atendimento de um cliente for exponencial, com um parâmetro fixo, pode-se obter, de imediato, três modelos por inversão das distribuições assumidas, para as chegadas e para os atendimentos, nos modelos M/G/1, M/D/1, M/E_k/1, obtendo então os modelos G/M/1, D/M/1 e E_k/M/1 [9].

O modelo G/M/1 não impõe qualquer restrição à distribuição associada ao processo de chegadas; o modelo D/M/1 assume chegadas em intervalos regulares e o modelo E_k/M/1 permite modelar um processo de chegadas que, não sendo com uma distribuição de Poisson, também não é determinístico (intervalos de tempo constantes).

2.4.10 Modelos com disciplinas de prioridade

Em certos sistemas de filas de espera o atendimento não é feito apenas por ordem de chegada, mas existe um sistema de prioridades, pelo que o atendimento de um cliente é feito pela respetiva prioridade.

Os modelos com prioridades são muito importantes, por exemplo, para as organizações de saúde. Enquanto o modelo base assume que os clientes são indistinguíveis e servidos pela disciplina FCFS, o modelo com prioridades assume que os clientes são distribuídos em duas ou mais classes de serviço *i = 1,2,…,N*, sendo servidos por ordem de prioridade, em que 1 é a mais alta e N é a mais baixa. Dentro da mesma classe, os clientes são servidos

pela disciplina FCFS. Mas quando há uma fila e um servidor fica disponível, um cliente pertencente a classe *i* só será servido se não houver nenhum cliente das classes *1,…,i-1* à espera [8].

Por exemplo, num serviço de urgências de uma unidade hospitalar, enquanto que poderá haver utentes que não agravam a sua situação clínica, caso tenham que esperar mais de uma hora para serem vistos pelo médico, poderá haver outros que são urgentes e que precisam de cuidados de um médico o mais rápido possível para evitar agravar a situação clínica. Neste caso, um modelo de filas com prioridade faz todo o sentido. O sistema de triagem de Manchester é um dos exemplos destes modelos mais usado nas urgências hospitalares.

O tratamento analítico de sistemas com prioridades é, obviamente, mais complicado do que o de sistemas sem prioridades. Como consequência, apenas se dispõe maioritariamente de resultados para o caso de um único servidor. Contudo, há um sistema com múltiplos servidores que apresenta resultados interessantes. Caracterizando-o [9]:

- Assume-se que existem N classes de prioridade (a classe 1 com prioridade mais elevada e a classe N com mais baixa prioridade). Os clientes são atendidos por ordem das suas classes de prioridade e, dentro da cada classe, por ordem de chegada;
- Assume-se que o processo de chegadas segue uma distribuição de Poisson, permitindo-se que a taxa de chegadas de clientes das várias classes possa ser diferente;
- Assume-se que as durações de atendimento são exponenciais para cada classe, considerando-se, adicionalmente, que a duração média de atendimento é igual para todas as classes.

De notar que, caso se ignorem as prioridades, estaremos perante o modelo M/M/s. Assim, quando se contabilizar o número total de clientes no sistema, pode considerar-se as distribuições limite apresentadas para o modelo M/M/s. Consequentemente, para um cliente selecionado aleatoriamente, são válidas as expressões obtidas para L_s, L_q, W_s e W_q nesse modelo. O que muda é a distribuição do tempo de espera. Neste caso, num modelo com prioridades, a variância da distribuição do tempo de espera aumenta, ou seja, teremos clientes de prioridade mais elevada com tempos de espera mais baixos do que ocorreriam com a disciplina FCFS sem prioridades. Por outro lado, os clientes de prioridade mais baixa terão tempos de espera mais elevados, o que não é de estranhar já que se pretende

melhorar o desempenho do sistema, no que diz respeito aos clientes de mais elevada prioridade, à custa de um pior desempenho para os clientes de mais baixa prioridade. Assim, é importante calcular o tempo de espera médio para um cliente de cada classe de prioridade.

As prioridades podem ser classificadas em duas categorias, as "não absolutas" (*nonpreemptive priorities*) e as "absolutas" (*preemptive priorities*) [1].

2.4.10.1 Prioridades "não absolutas" (*nonpreemptive priorities*)

São aquelas em que um cliente que está a ser atendido não vê o seu atendimento interrompido pela chegada de um cliente com prioridade mais elevada. Após cada conclusão do serviço, o próximo cliente a entrar é escolhido pela ordem de prioridade, baseado numa ordem FCFS dentro da mesma classe [24].

Considerando as prioridades "não absolutas", W_k, será o tempo de espera médio para um cliente da classe de prioridade k (incluindo a duração do atendimento) e será dado por:

$$W_k = \frac{1}{A.B_{k-1}.B_k} + \frac{1}{\mu}; \text{para } k = 1,2,\dots,N \text{ (N-classes de prioridade)}$$

onde

$$A = s!\left(\frac{s\mu - \lambda}{\rho^s}\right)\sum_{j=0}^{s-1}\frac{\rho^j}{j!} + s.\mu\,; \quad B_0 = 1;\ B_k = 1 - \frac{\sum_{i=1}^{k}\lambda_i}{s.\mu}, \text{para } k = 1,2,\dots,N$$

sendo

s = número de servidores;

μ = taxa média de serviço por cada servidor ocupado;

λ_i = taxa média de chegadas da classe de prioridade i, i = 1,2,…,N

$$\lambda = \sum_{i=1}^{N}\lambda_i; \quad \rho = \frac{\lambda}{\mu}$$

Estes resultados assumem que $\sum_{i=1}^{k}\lambda_i < s.\mu$, de modo a que a classe de prioridade k possa atingir um estado de equilíbrio.

Para cada classe de prioridade aplica-se a Fórmula de Little (subsecção 2.3), pelo que o número esperado de clientes da classe de prioridade *k* no sistema, incluindo os que estão a ser atendidos, será

$$L_k = \lambda_k.W_k, \text{ para } k = 1,2,\ldots,N.$$

Para estes tipos de sistemas também são válidas as fórmulas seguintes:

- Tempo médio de espera, para a classe k, a aguardar atendimento:

$$W_k - (1/\mu); \text{ para } k = 1,2,\ldots,N$$

- Comprimento médio da fila de espera correspondente à classe k:

$$\lambda_k.[(W_k - (1/\mu)]; \text{ para } k = 1,2,\ldots,N$$

- Se $s = 1$, então $A = \mu^2/\lambda$.

Na notação de Kendall, este tipo de prioridade é indicada como NPRP (*Nonpreemptive Priority Models*) na característica da disciplina da fila [24].

2.4.10.2 Prioridades "absolutas" (*preemptive priorities*)

São aquelas em que o atendimento de um cliente será interrompido e reenviado para a fila de espera, devido à chegada de um cliente com mais elevada prioridade. O serviço do cliente anterior continuará a partir do ponto em que foi interrompido, depois do cliente prioritário estar servido [24].

Mantendo as hipóteses já referidas, W_k corresponde ao tempo de espera médio para um cliente da classe de prioridade *k* (incluindo a duração do atendimento) e será, para um único servidor, dado por:

$$W_k = \frac{1/\mu}{B_{k-1}.B_k}; \text{para } k = 1,2,\ldots,N$$

O número esperado de clientes da classe de prioridade *k* no sistema será:

$$L_k=\lambda_k.W_k; \text{ para } k = 1,2,\ldots,N$$

Refira-se, finalmente, que dado que a distribuição exponencial não tem memória, as interrupções de atendimento não afetam, em média, o processo de atendimento, ou seja, a duração média total do atendimento continua a ser igual a $1/\mu$. Quando um cliente com atendimento interrompido voltar a ser atendido, a distribuição da duração do atendimento restante continuará a mesma [9].

Na notação de Kendall, este tipo de prioridade é indicada como PRP (*Preemptive Priority Models*) na característica da disciplina da fila [24].

Por razões óbvias, a disciplina de prioridades "absolutas", raramente são usadas caso os elementos que constituem a fila de espera sejam pessoas [24].

2.5. Redes de filas de espera

Neste estudo, até agora, tem-se considerado, sobretudo, sistemas de filas de espera com um único local de atendimento, ainda que com um ou mais servidores. No entanto, em muitas situações reais, como por exemplo, num serviço de urgências de uma unidade hospitalar, onde um utente terá de passar por uma sequência de filas de espera, seguindo, ou não, uma determinada ordem, haverá um sistema com múltiplas fases (Figura 2.9 e Figura 2.11). Estas fases são compostas pela inscrição inicial, pela triagem, pela consulta, pela realização de exames complementares de diagnóstico, etc. O *output* de algumas dessas filas será o *input* de outras. Estaremos, assim, perante um sistema de redes de filas de espera. Quando tal ocorre, é importante estudar globalmente a rede para determinar, para além dos tempos de espera ou o número de clientes por fase, também o tempo total de espera ou o número total de clientes no sistema.

Assim, no caso concreto deste estudo, cuja explicação será aprofundada nos capítulos seguintes, teremos, no serviço de urgências, um sistema de fila de espera para o registo da admissão inicial, cujo *output* será o *input* do segundo sistema de fila de espera – a triagem, que por sua vez terá um *output* que corresponderá ao *input* do terceiro sistema de fila de espera – a consulta. Portanto, teremos um atendimento que é constituído por diversas tarefas, cada uma das quais dando origem a uma fila, isto é, estaremos na presença de um sistema em rede de filas de espera. Será com base neste pressuposto que se irá modelar o sistema, tal como se explica nos capítulos seguintes.

2.6. Considerações finais sobre as filas de espera

A importância das filas de espera é notória no dia-a-dia e em variados contextos. Assim, é evidente que a gestão adequada de um sistema de filas de espera tem repercussões na qualidade de vida e na produtividade.

Na modelação matemática de sistemas de filas de espera, a distribuição exponencial tem um papel fulcral para representar as distribuições dos tempos de chegadas e os tempos de serviço, ainda que em determinadas situações possa ser útil considerar outras distribuições, nomeadamente a distribuição de Erlang.

De referir ainda a necessidade de, em certos sistemas, se tornar necessário separar os clientes em diferentes classes, cada uma das quais com um nível de prioridade distinto.

Quando um cliente precisa de recorrer a vários serviços, num mesmo sistema, torna-se útil modelá-lo como uma rede de filas de espera.

Refira-se, finalmente, que quando há particularidades especiais, não contempladas em qualquer modelo conhecido, poderemos recorrer à simulação de filas de espera, através de um programa de computador, para simular o funcionamento do sistema. Esta técnica será objeto de desenvolvimento, através de um caso de estudo, nos capítulos seguintes.

3. Metodologia

Depois de obter a base de dados, fornecida pelo Hospital de Santo André de Leiria, efetuou-se uma análise geral e constatou-se que a base de dados necessitava de alguns ajustes que se descrevem nos tópicos seguintes. Posteriormente, estes ajustes permitem analisar a base de dados através do *software* R [11, 22].

Composta por 91206 registos mais o cabeçalho das colunas, a base de dados, formada por um ficheiro Excel (xlsx) constituído por duas folhas, contém os dados relativos aos utentes que deram entrada no serviço de urgência durante o ano de 2014.

É constituída pelas variáveis seguintes referidas pelo nome e ordem original da base de dados: data nascimento, sexo, prioridade, data admissão, TriagemUG Leiria, data_triagem, primeira_observa_med, data da observação, pm_manchester, data alta, destino.

3.1. Descrição das variáveis

- "data nascimento" – contém a data de nascimento do utente no formato dd/mm/aaaa.
- "sexo" – contém o sexo do utente, representado por F (feminino) e M (masculino).
- "prioridade" – contém um número de 1 a 5 e uma cor segundo a classificação da Triagem de Manchester.
- "data admissão" – contém a data e a hora que representa o tempo registado correspondente ao momento que o funcionário recebe o utente de modo a efetuar a sua inscrição na admissão inicial.
- "TriagemUG Leiria" – designação do nome da triagem do serviço de urgências do hospital de Leiria.
- "data_triagem" – contém a data e a hora que representa o tempo registado correspondente ao momento que o enfermeiro recebe o utente quando este chega à triagem.

- “primeira_observa_med” – descreve a primeira observação feita pela triagem, nomeadamente qual a especialidade médica atribuída ao utente.

- “data da observação” – contém a data e a hora que representa o tempo registado correspondente ao momento que o médico recebe o utente quando este chega à consulta.

- “pm_machester” – descreve os sintomas identificados pela Triagem de Manchester.

- “data alta” – contém a data e a hora que representa o tempo registado correspondente ao momento em que o utente abandona o hospital.

- “destino” – representa o destino dado ao utente, nomeadamente médico de família, internamento, alta médica, etc.

3.2. Alterações efetuadas na base de dados

As alterações foram efetuadas, pela ordem seguinte, no ficheiro Excel da base de dados (formato .xlsx) composta por 91206 registos mais o cabeçalho das colunas distribuídos por duas folhas, tal como referido anteriormente.

Eliminaram-se, nas duas folhas, as colunas J e L que estavam vazias.

Na folha 1, colocaram-se as horas e as datas das variáveis (colunas) “data admissão”, “data_triagem”, “data da observação” e “data alta”, em células separadas, uma vez que, inicialmente, data e hora estavam na mesma célula. A hora foi retirada da data e colocada numa coluna nova através do menu Dados - Texto para colunas.

Foi efetuado o mesmo procedimento na folha 2. A base de dados ficou com 4 novas variáveis que são: “hora admissão”, “hora triagem”, “hora da observação” e “hora alta”.

Eliminou-se a coluna com a variável “TriagemUG Leiria”, nas duas folhas, por ser sempre igual e não ser relevante para a análise dos dados.

Acrescentaram-se os dados da folha 2 à folha 1 de modo a que a base de dados fosse constituída por uma só folha e eliminou-se a folha 2.

Ordenou-se a coluna com a variável "data admissão" e depois a coluna com a variável "hora admissão" por ordem crescente em ambas.

De modo a facilitar a análise através do *software* R, atribuíram-se abreviaturas aos títulos das colunas passando a base de dados a ser formada pelas variáveis com as respetivas abreviaturas, conforme a Tabela 3.1.

Variáveis da Base de Dados	
Designação original	**Nova designação (abreviatura)**
data nascimento	dn
sexo	sexo
data admissão	da
hora admissão	ha
data_triagem	dt
hora triagem	ht
prioridade	prio
primeira_observa_med	pom
data da observação (consulta)	do
hora da observação (consulta)	ho
pm_manchester	pm
data alta	dal
hora alta	hal
destino	dest

Tabela 3.1 - Variáveis da Base de Dados

De seguida efetuaram-se alguns cálculos que permitirão uma análise mais detalhada dos dados.

Calculou-se *Δha = (hora admissão do utente) – (hora admissão do utente anterior)*. Esta variável, *Δha*, poderia dar-nos o tempo da duração do serviço de admissão, ou seja, o tempo que cada utente demorou a ser atendido, permitindo assim obter a taxa de serviço do sistema de admissão inicial (μ_a), admitindo que o serviço de admissão não está vazio quando há utentes em espera. Como a base de dados não fornece todos os dados, a variável *Δha*, inclui o tempo de registo na admissão inicial, mas igualmente o tempo no qual nenhum utente aguardava o registo, havendo ainda a complexidade de existirem 2 funcionários para o registo da admissão.

Calculou-se *tet = ht - ha (hora triagem - hora admissão)*, relativamente ao mesmo utente. Esta variável, *tet*, poderia dar-nos o tempo de espera para a triagem, caso se soubesse a

hora de saída da admissão, dados que não são fornecidos pela base de dados. Assim sendo, esta variável representa o tempo de admissão adicionado ao tempo de espera para a triagem.

Calculou-se *Δht = (tempo de chegada à triagem do utente) – (tempo de chegada à triagem do utente anterior)*. Esta variável, *Δht*, poderia dar-nos o tempo da duração do serviço de triagem, ou seja, o tempo que cada utente demorou na triagem, permitindo assim obter a taxa de serviço do sistema de triagem (μ_t), admitindo que o serviço de triagem não está vazio quando há utentes em espera ou caso se soubesse a hora que o utente saiu do serviço de triagem. Como estes dados não são fornecidos pela base de dados, *Δht* inclui o tempo do serviço de triagem mais o tempo no qual ninguém está na fila de espera da triagem.

Depois de efetuar este cálculo verificou-se que havia registos de utentes que chegavam antes de outros à admissão e eram atendidos depois deles na triagem, ou seja, a hora da triagem do utente seguinte era inferior à do anterior, fazendo com que houvesse horas negativas na duração do serviço da triagem. Para evitar este tipo de situação, ordenaram-se as colunas data e hora da triagem e, em seguida, calculou-se *Δht*. Depois, colaram-se apenas os valores obtidos numa nova coluna e ordenaram-se as colunas data e hora de admissão, tal como inicialmente. Por último, eliminou-se a coluna de *Δht* que continha as fórmulas.

Calculou-se *tec = ho - ht (hora observação – hora triagem)*, relativamente ao mesmo utente. Caso se soubesse a hora de saída da triagem, dados que não são fornecidos pela base de dados, esta variável, *tec*, poderia dar-nos o tempo de espera para a consulta (observação). Por conseguinte, a variável inclui o tempo de espera para a observação mais o tempo de serviço da triagem.

Calculou-se *tsc = hal-ho (hora alta - hora observação)*, relativamente ao mesmo utente. Esta variável, *tsc*, pode dar-nos o tempo da duração do serviço de consulta, ou seja, o tempo que cada utente demorou na consulta, permitindo assim obter a taxa de serviço do sistema de consulta (μ_c), caso se considere que a hora de alta coincide com a hora de saída da consulta. Todavia, tal análise não foi efetuada, pois efetivamente este tempo inclui diversas situações bem distintas, tais como o tempo de consulta ou o tempo no qual o utente fica internado devido a uma intervenção cirúrgica.

Depois destes cálculos, verificou-se que a base de dados continha algumas incongruências que foram eliminadas.

Eliminaram-se registos onde a hora de alta era inferior à hora de observação e registos com valores de tempos de espera, da admissão para a triagem, demasiadamente altos que aconteciam apenas quando a entrada do utente correspondia à hora 00:00:00, o que se depreendeu serem erros nos registos.

Eliminaram-se também registos com tempos de espera, da triagem para a consulta, superiores a 02:30:00 por parecerem valores desajustados.

Eliminaram-se ainda 905 registos repetidos que continham a mesma data de nascimento e tempos iguais, 749 registos que continham prioridade 6 (cor branca) e que não consta no sistema de Triagem de Manchester e 198 registos onde a hora de triagem era inferior à hora de admissão.

Por último, corrigiram-se 38 registos onde a data de alta era inferior à data de observação.

Com estes ajustes eliminaram-se 4033 registos, que correspondem a 4,4% do total de registos iniciais, ficando a base de dados final com 87173 registos.

Por haver utentes com entrada no final do dia 31 de dezembro, a saída foi realizada já no dia 1 de janeiro, não estando estes dados registados na base de dados, o que originou células vazias. Por isso, e de modo a que estes dados também sejam interpretados pelo *software* R, foi colocado, nestas células, o valor NA (*Not Available*).

Para finalizar o tratamento da base de dados, colocaram-se as colunas das variáveis com uma ordem lógica, ficando a coluna das prioridades e dos primeiros sintomas obtidos pela triagem a seguir à coluna dos tempos da triagem.

Deste modo, a base de dados final, e depois de efetuados os cálculos referidos anteriormente, ficou com as variáveis referidas na Tabela 3.2 e pela ordem indicada, estando a primeira variável na coluna 1 da base de dados, a segunda variável na coluna 2 e assim sucessivamente.

Tendo em conta que a importação da base de dados pelo software R será feita através de um ficheiro com formato .csv, indicam-se, na Tabela 3.2, os nomes das variáveis finais

depois de convertidas, pelo Excel, para um ficheiro com formato .csv.

Variáveis finais da Base de Dados	
Descrição	**Nome da variável**
Data de nascimento	dn
Sexo	sexo
Data de admissão	da
Hora de admissão	ha
Δha = (hora admissão do utente) – (hora admissão do utente anterior)	Dha
Data da triagem	dt
Hora da triagem	ht
tet = ht – ha	tet
Δht = (tempo de chegada à triagem do utente) – (tempo de chegada à triagem do utente anterior)	Dht
Prioridade	prio
Primeira observação médica	pom
Pm manchester	pm
Data da observação	do
Hora da observação	ho
tec = ho – ht	tec
Data da alta	dal
Hora da alta	hal
tsc = hal-ho	tsc
Destino	dest

Tabela 3.2 - Nomes das variáveis depois de convertidas para o formato .csv

3.3. Resultados obtidos da base de dados

Alguns resultados relevantes, obtidos da base de dados, são mostrados nas figuras seguintes.

A Figura 3.1 mostra o número de utentes, que chegaram, por mês, às urgências, ao longo do ano de 2014, onde se pode constatar que o mês com maior afluência foi o de agosto, com 8105 utentes, e o mês com menor número de utentes foi o de fevereiro com 6527.

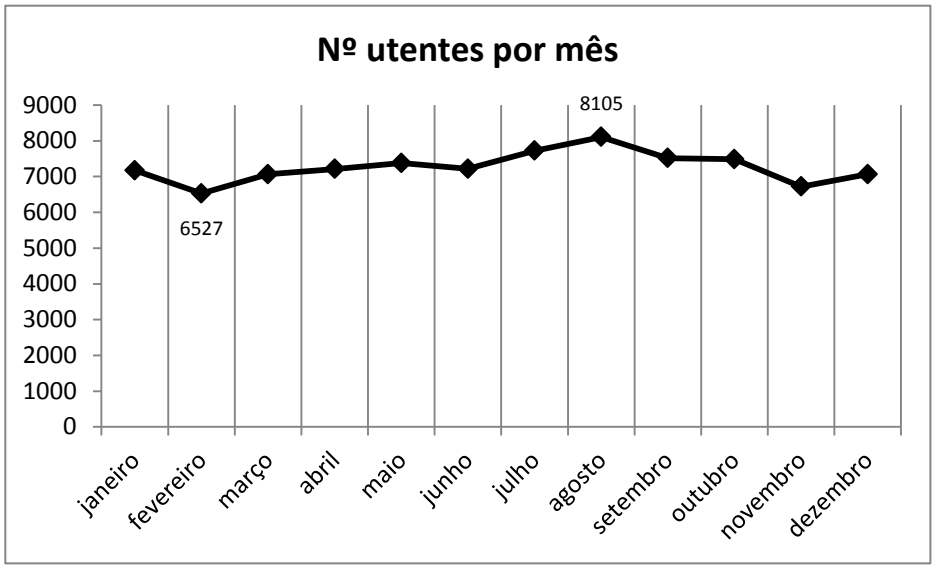


Figura 3.1 - Número de utentes por mês

Como nem todos os meses têm o mesmo número de dias representa-se, na Figura 3.2, a média diária dos utentes que chegaram às urgências, em cada mês. Verifica-se, novamente, que o mês com maior afluência foi o de agosto, com uma média de 261,5 utentes por dia. No entanto, o mês com menos afluência de utentes por dia, passa a ser o de novembro com uma média diária de 224 utentes.

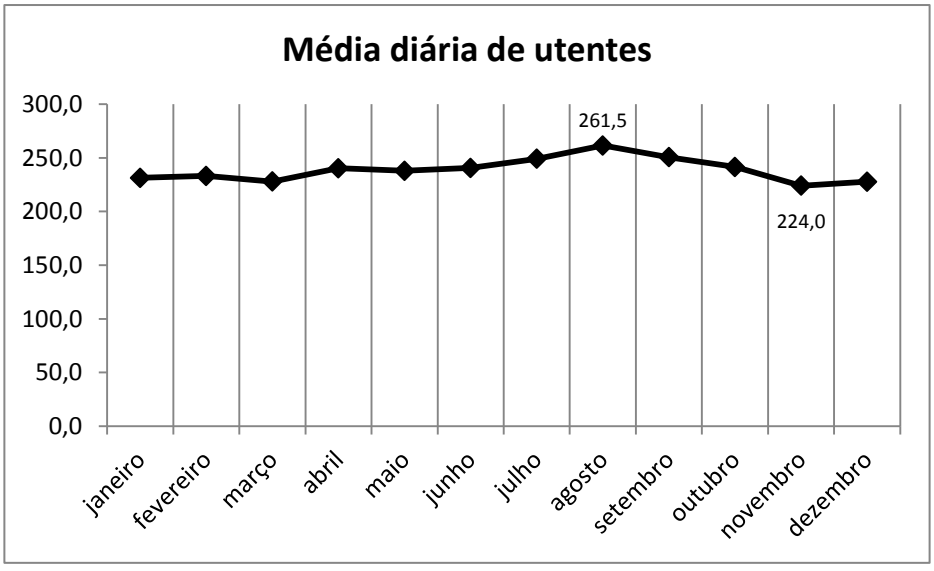


Figura 3.2 - Média diária de utentes por mês

A Figura 3.3 mostra a afluência de utentes, às urgências, por dia de semana. O dia com mais utentes é a segunda-feira com 14402 e o dia com menos utentes é o domingo com 10677. Estes valores são totais, ou seja, o número de utentes em cada dia, por exemplo ao domingo, representa a soma dos utentes que chegaram em todos os domingos de 2014.

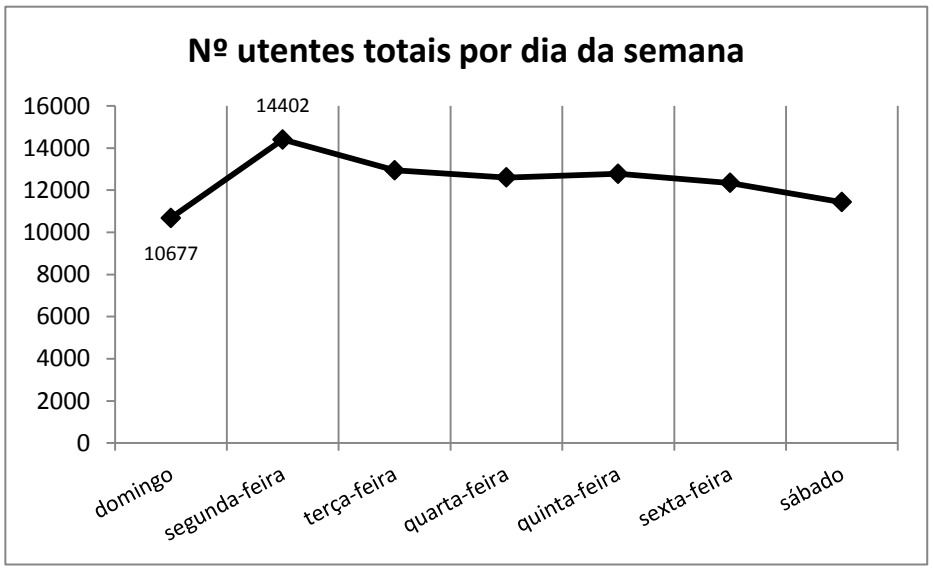


Figura 3.3 - Número de utentes totais por dia da semana

A Figura 3.4 mostra a afluência de utentes, às urgências, em cada hora do dia. A hora com mais utentes é das 10h às 11h com 7108 e a hora com menos afluência é das 4h às 5h com 915 utentes. Estes valores são totais, ou seja, o número de utentes representa a soma de todos os utentes que chegaram nessa hora ao longo do ano.

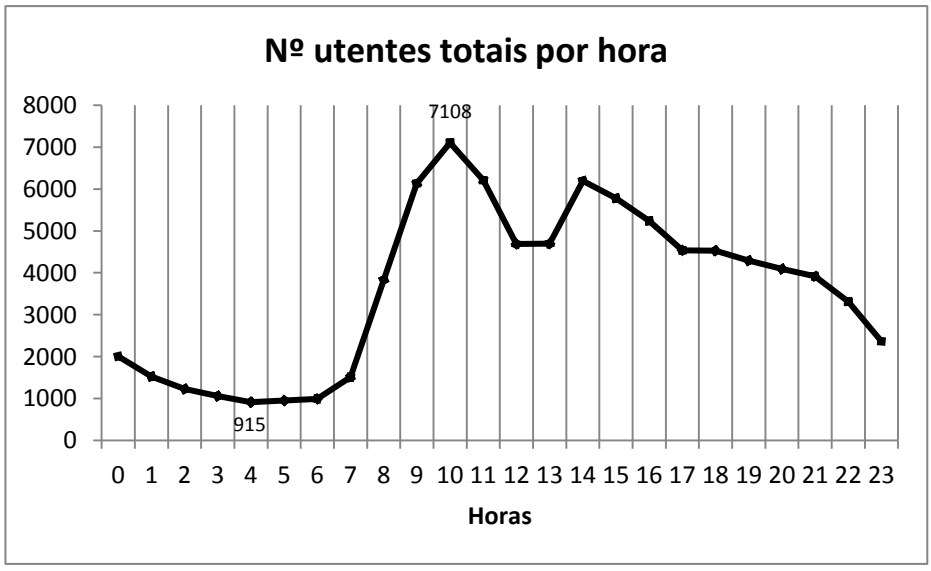


Figura 3.4 - Número de utentes totais por hora

A Figura 3.5 mostra o número de utentes, por hora, em dias específicos. Escolheu-se o domingo e a segunda-feira por serem os dias da semana com menos e mais registos de utentes, respetivamente, e os meses de novembro e agosto por serem os meses com menos e mais afluência de utentes, respetivamente.

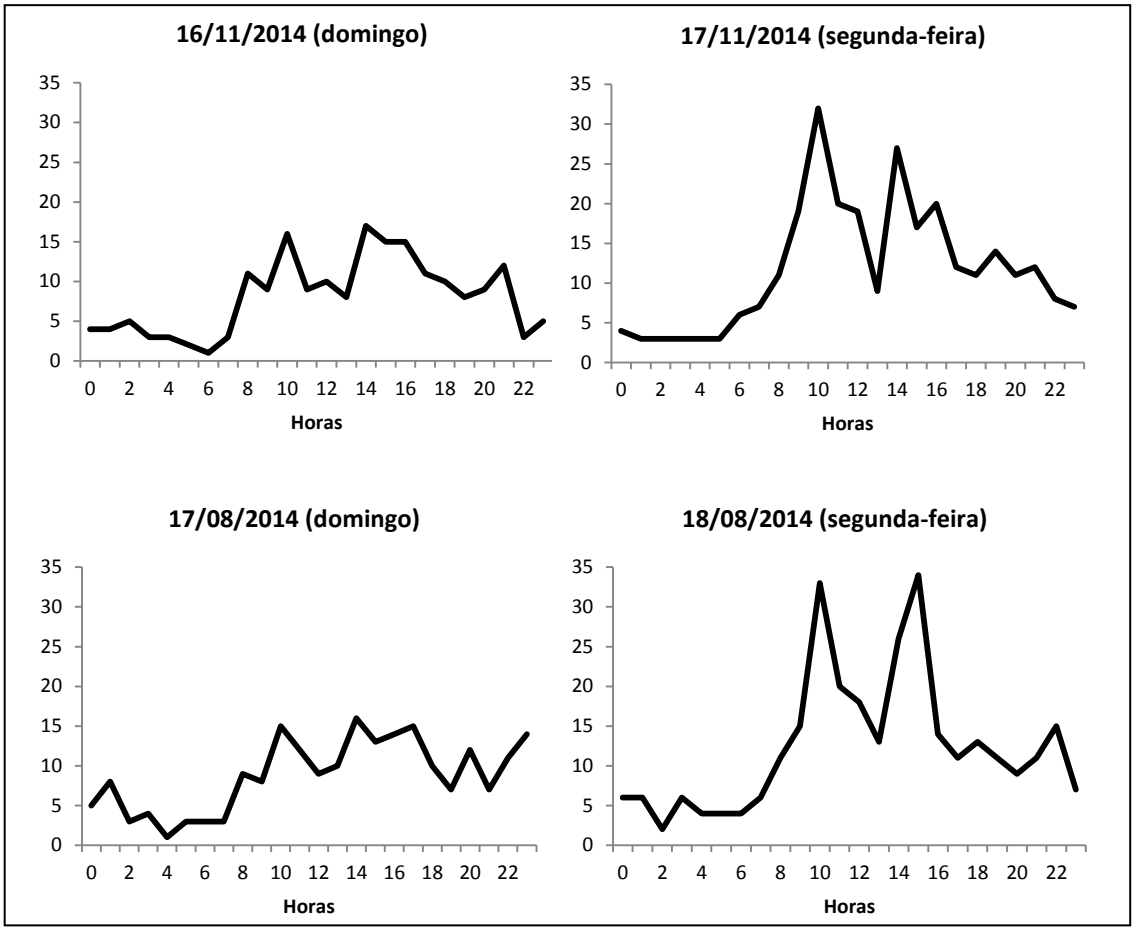


Figura 3.5 - Número de utentes em dias específicos

Por último, a Figura 3.6 mostra a prioridade atribuída aos utentes depois da Triagem de Manchester. Depois da triagem, verifica-se que a prioridade 3 (cor amarela com situação urgente) é a que tem o maior número de utentes e a prioridade 1 (cor vermelha com situação emergente) a que tem menos utentes ao longo do ano.

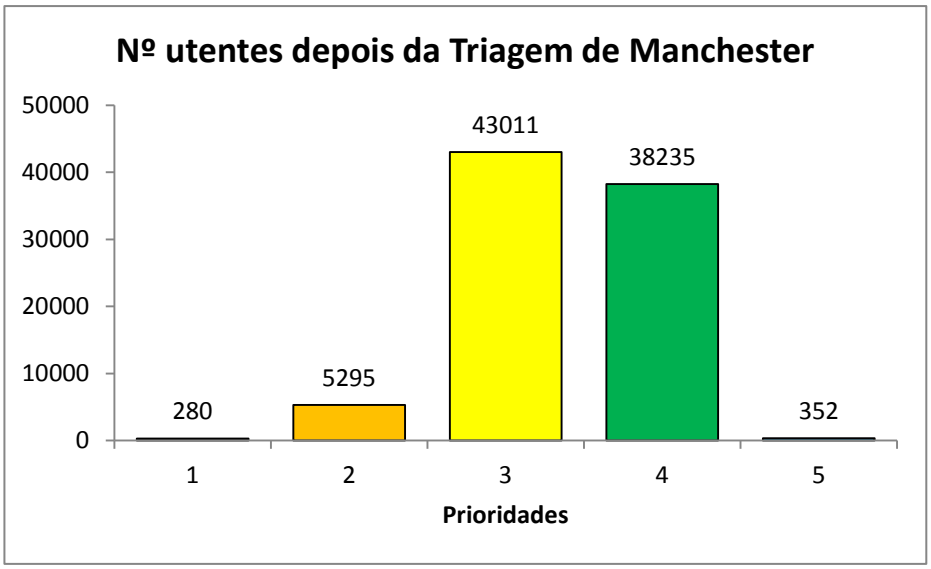


Figura 3.6 - Número de utentes pela Triagem de Manchester

3.4. Simulação

Para ser realizada uma simulação, é necessário construir um modelo computacional que corresponda à situação real que se deseja simular, permitindo obter uma descrição aproximada das características do sistema em estudo.

Assim, utilizando o *software* R [11, 22], a programação do modelo foi feita tentando aproximar o mais possível os dados da simulação aos dados da base de dados. Para isso, definiram-se os parâmetros correspondentes às taxas média de chegada, taxas média de serviço e o número de servidores em cada uma das fases do sistema de filas de espera.

Considerando que o sistema em estudo é do tipo múltiplos servidores com múltiplas fases (Figura 2.11) e tendo em conta as informações fornecidas pelo hospital, as variáveis foram definidas seguindo os seguintes critérios:

- No serviço de admissão do registo inicial do utente, o número de servidores está fixo durante todo o ano, havendo 2 funcionários (2 servidores) das 08:00h às 00:00h e 1 funcionário (1 servidor) das 00:00h às 08:00h.

- No serviço de triagem, o número de enfermeiros é variável havendo sempre 1 enfermeiro (1 servidor) fixo, mas, por vezes, pode ser necessário aumentar este número. No entanto, esta informação, sobre quando é que há necessidade de aumentar

o número de enfermeiros na triagem e de quanto pode ser esse aumento, não foi fornecida pelo hospital.

- No serviço de consulta, o número de médicos (servidores) é variável. Nas consultas de especialidade existem escalas fixas de especialistas 24h por dia, formadas por 3 ortopedistas, 3 cirurgiões, 3 internistas e ainda 1 psiquiatra das 08:00h às 20:00h. Quanto aos médicos generalistas existem 3 das 08:00h às 14:00h, 4 das 14:00h às 00:00h e 2 das 00:00h às 08:00h, podendo, nalguns casos, existir alterações no número de médicos generalistas.

Face aos critérios anteriores, e devido à complexidade e a alguma falta de informação, optou-se por implementar um sistema de fila de espera com três fases: admissão, triagem e consulta. Na fase de admissão, consideraram-se 2 servidores, na triagem 1 servidor e na fase consulta 4 servidores. Em todas as fases o sistema de fila é do tipo fila única com uma disciplina do tipo FCFS (*First Come First Served*), com exceção da fila para as consultas, uma vez que, após a triagem, são definidas as prioridades através do sistema de Manchester. O sistema geral é mostrado na Figura 3.7.

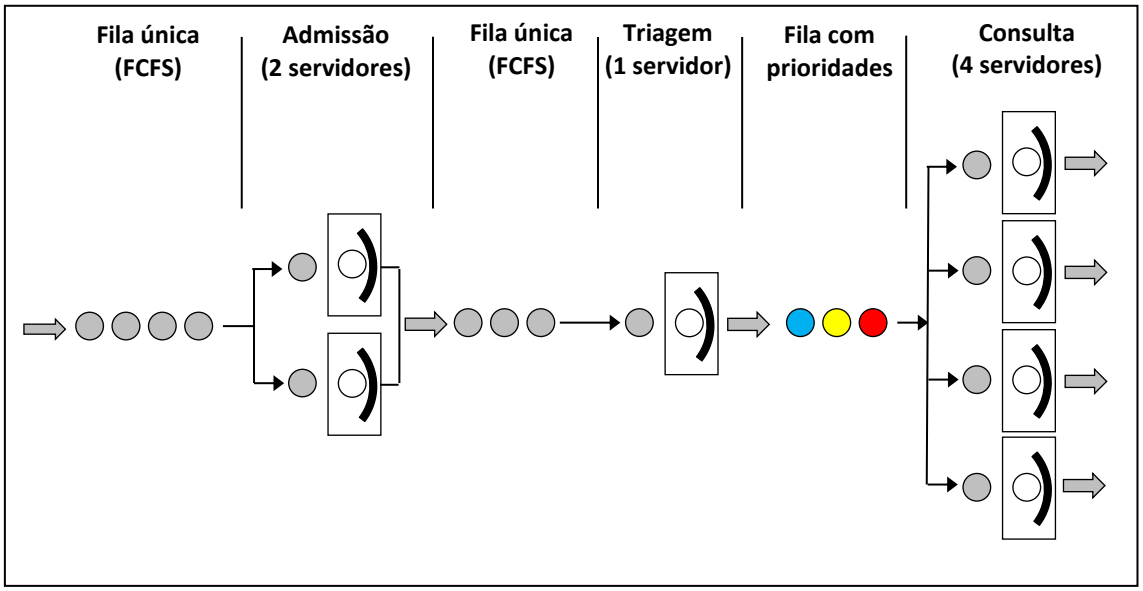


Figura 3.7 - Sistema geral da fila de espera em estudo

As variáveis e os parâmetros, utilizados na programação da simulação, foram definidos conforme a Tabela 3.3, onde as taxas de chegada e de serviço correspondem a valores por hora. Os valores dos parâmetros seriam válidos caso se usasse a base de dados completa, mas, como se explica na subsecção 3.4.1, estes valores serão alterados devido ao facto de se usar um subconjunto da base de dados.

Variáveis e parâmetros da simulação com base de dados completa			
Fase	**Descrição**	**Nome**	**Valor**
Admissão	Taxa média de chegada dos clientes à admissão	Tc	10
	Taxa média de serviço de um servidor da admissão	Ts	10
	N.º de servidores na admissão	ns	2
Triagem	N.º de utentes que vão entrar na fila de espera da triagem	nt	variável
	N.º de utentes que chegaram à triagem	nut	variável
	Taxa média de serviço de um servidor da triagem	Tst	10
	N.º de servidores na triagem	nst	1
	Definição das prioridades da triagem	triagem	variável
	Probabilidade de atribuição da cor vermelha após triagem	p1	0.00321
	Probabilidade de atribuição da cor laranja após triagem	p2	0.06074
	Probabilidade de atribuição da cor amarela após triagem	p3	0.49339
	Probabilidade de atribuição da cor verde após triagem	p4	0.43861
	Probabilidade de atribuição da cor azul após triagem	p5	0.00404
Consulta	N.º de utentes que chegaram à consulta	nuc	variável
	Taxa média de serviço de um servidor da consulta	Tsc	4
	N.º de servidores nas consultas	nsc	4
Notação comum às 3 fases	N.º de utentes que chegaram ao hospital	nu	variável
	N.º de utentes nas diversas fases do sistema	n	variável
	Tempo da próxima chegada às diversas fases do sistema	t1	variável
	Tempo da próxima chegada	t2	variável
	Tempo final da simulação	t.end	1000
	Relógio	t.clock	variável
	Matriz de dados	T	variável
	N.º de linhas da matriz T	dimensao	variável

Tabela 3.3 - Variáveis e parâmetros utilizados na simulação

O parâmetro *Tc = 10* foi calculado dividindo o número total de registos da base de dados pelo número de horas do ano (*87173/(365*24)≈10*), obtendo-se, assim, o número médio de utentes que chegam à admissão, por hora. O parâmetro *Ts = 10* foi obtido calculando a média do tempo da duração do serviço na admissão, cujo resultado foi de 6.03 minutos. Em seguida, dividiu-se uma hora pela média obtida (*60/6.03≈10*), obtendo-se, assim, o número de utentes atendidos por hora. Os restantes parâmetros das taxas médias de serviço foram obtidos de forma similar. Os parâmetros das probabilidades de atribuição da cor após a triagem (*p1,…,p5*) foram obtidos através do comando *prop.table(table(h$prio))*, que devolve uma tabela com as proporções da variável *prio*, existente na base de dados.

A simulação consiste em gerar números aleatórios que constroem, através da função *rbind*, uma matriz T com uma estrutura de 8 colunas e onde cada linha corresponde aos dados que caracterizam cada utente. Tendo em conta a estabilidade da simulação, na matriz final, foram retirados os primeiros valores que poderiam causar alguma instabilidade. O significado de cada coluna, da matriz, está explicado na Tabela 3.4.

Colunas da matriz T	
Nº da coluna	**Descrição**
1	Hora de chegada ao hospital
2	Hora de início da admissão
3	Hora de fim da admissão
4	Hora de início da triagem
5	Hora do final da triagem
6	Resultado da triagem
7	Hora de entrada na consulta
8	Hora de saída da consulta

Tabela 3.4 - Estrutura da matriz T

Os números gerados relacionados com a hora de fim da admissão (coluna 3) correspondem à hora de início da fila de espera para a triagem. Assim, os dados da coluna 3 serão iguais aos da coluna 4 (hora de início da triagem) quando o servidor da triagem estiver livre, uma vez que, neste caso, não há fila de espera nesta fase. Isto acontece em termos teóricos, mas, na realidade, haverá sempre um tempo de deslocação entre os postos.

Os números gerados para a hora do final da triagem (coluna 5) correspondem à hora de início da fila de espera para a consulta. Assim, os dados da coluna 5 serão iguais aos da coluna 7 (hora de entrada na consulta) quando um dos 4 servidores da consulta estiver livre, uma vez que, nesta situação, não há fila de espera nesta fase. Tal como no caso anterior, este facto é válido em termos teóricos, uma vez que, na realidade, haverá sempre um tempo de deslocação entre os postos.

O código completo do programa da simulação é apresentado no Anexo I.

Um exemplo de uma parte da matriz T, com 10 utentes ordenados pela hora de chegada ao hospital, gerada pelo *software* R, é mostrado na Figura 3.8.

```
> T
          [,1]     [,2]     [,3]     [,4]     [,5] [,6]     [,7]     [,8]
  ⋮         ⋮        ⋮        ⋮        ⋮        ⋮    ⋮        ⋮        ⋮
[141,] 205.5305 206.6931 207.5058 207.5058 207.5334    4 207.7062 209.4643
[142,] 206.3429 206.7707 207.2645 207.2645 207.6211    3 207.6782 209.3311
[143,] 206.4176 207.2645 207.7954 207.7954 208.6556    3 208.6556 209.4316
[144,] 207.1583 207.5058 207.9745 207.9745 208.1252    4 208.2681 210.3923
[145,] 207.4389 207.7954 208.5910 208.5910 209.3673    3 209.4316 209.9548
[146,] 207.8972 207.9745 208.8892 208.8892 208.9786    4 209.4643 211.9667
[147,] 208.0778 208.5910 208.9532 208.9786 209.0377    2 209.3311 210.3938
[148,] 208.2460 208.8892 209.1143 209.1143 211.1090    4 211.1090 212.3957
[149,] 208.9959 208.9959 209.8463 210.0085 210.4603    4 210.4603 210.4691
[150,] 209.1371 209.1371 209.5067 209.5067 210.0085    3 210.0085 210.6588
  ⋮         ⋮        ⋮        ⋮        ⋮        ⋮    ⋮        ⋮        ⋮
```

Figura 3.8 - Matriz T parcial simulada pelo software R

Analisando a matriz da figura anterior e considerando que os números gerados representam tempos (por exemplo, minutos), pode-se observar que, por exemplo, o utente número 147 chega ao hospital no minuto 208.0778 (coluna 1) e está em fila de espera até que comece a ser atendido na admissão inicial, no minuto 208.5910 (coluna 2). Termina o seu registo no minuto 208.9532 (coluna 3) e espera na fila até que comece a ser atendido na triagem, o que acontece no minuto 208.9786 (coluna 4), terminando a triagem no minuto 209.0377 (coluna 5). O resultado da triagem para este utente foi o 2 (coluna 6), que corresponde à cor laranja do sistema de Manchester. Depois da triagem, este utente, e após um tempo em fila de espera, entra na consulta no minuto 209.3311 (coluna 7) e sai da consulta ao minuto 210.3938 (coluna 8). De referir também que este utente, pelo facto de ter uma prioridade mais urgente do que a do utente anterior (linha 146), entra na consulta primeiro, mesmo tendo saído da triagem depois desse utente.

Noutro exemplo, pode-se observar que o utente número 150 sai da triagem no minuto 210.0085 (coluna 5) e entra na consulta no minuto 210.0085 (coluna 7) que, por ser exatamente o mesmo momento no qual saiu da triagem, significa que não houve fila de espera entre a triagem e a consulta, ou seja, um servidor estava livre.

3.4.1 Comparação das distribuições da hora de início da admissão

De modo a que se consiga analisar os resultados e retirar conclusões dos dados fornecidos pela simulação é necessário que estes dados tenham a mesma distribuição estatística que os

dados da base de dados real. Por conseguinte, foi necessário comparar as duas amostras na expetativa que seguissem a mesma distribuição.

Para a comparação de duas amostras independentes, em que não se conhecem quais as suas distribuições, utilizam-se testes estatísticos não-paramétricos, que são testes que não especificam condições sobre os parâmetros nem sobre a distribuição uma vez que apenas comparam as observações das duas amostras.

Para o caso em estudo, o teste adequado é o *Teste de Kolmogorov-Smirnov* [2, 12, 23] que irá ter como hipóteses a testar:

H_0: as duas amostras são provenientes da mesma distribuição;

H_1: as duas amostras não são provenientes da mesma distribuição.

Pelo facto da base de dados do hospital ter apenas a hora de chegada à admissão inicial e não a hora de chegada ao hospital, optou-se por comparar estes dados com a coluna 2 da matriz T de simulação, que representa, tal como a base de dados, a hora de início da admissão.

De referir que enquanto na base de dados existem comportamentos bem diferenciados, por exemplo se considerarmos diferentes horas ou diferentes dias ao longo do ano, na matriz de simulação todos os dados são homogéneos, não havendo grandes diferenças entre os dados gerados dos diversos utentes. Por exemplo, conforme foi retratado anteriormente, o número de utentes entre as 03:00h e as 05:00h é bem distinto do número de utentes entre as 12:00h e as 14:00h, razão pela qual a base de dados referente ao primeiro período será caracterizada por uma distribuição totalmente distinta da que caracteriza o segundo período referido. Por este facto, para a comparação das distribuições entre os dados reais e os simulados, escolheu-se um subconjunto da base de dados que corresponda a um comportamento homogéneo. Considerando os gráficos da Figura 3.3 e da Figura 3.4, optou-se por selecionar os registos das quartas-feiras e quintas-feiras, entre as 14:30h e 16:30h, durante todos os dias do ano, fazendo com que a base de dados parcial ficasse com 3459 registos. Qualquer outro horário, desde que homogéneo, pode ser modelado de forma análoga à que iremos aplicar.

Tendo em conta o subconjunto selecionado, alguns parâmetros descritos na Tabela 3.3 foram recalculados. Assim, aplicando os mesmos métodos descritos na secção 3.4 e de modo a aproximar as duas distribuições o mais possível, chegou-se aos valores descritos na

Tabela 3.5 (alterações assinaladas com cor). Foi com estes valores que se efetuaram os testes de comparação das distribuições e, doravante, de cada vez que se fizer referência à base de dados entenda-se que esta engloba, apenas, o subconjunto atrás referido.

Parâmetros da simulação com o subconjunto da base de dados		
Descrição	**Nome da variável**	**Valor**
Taxa média de chegada dos clientes à admissão	Tc	15
Taxa média de serviço de um servidor da admissão	Ts	10
N.º de servidores na admissão	ns	2
Taxa média de serviço de um servidor da triagem	Tst	21
N.º de servidores na triagem	nst	1
Probabilidade de atribuição da cor vermelha após triagem	p1	0.00145
Probabilidade de atribuição da cor laranja após triagem	p2	0.05984
Probabilidade de atribuição da cor amarela após triagem	p3	0.52472
Probabilidade de atribuição da cor verde após triagem	p4	0.41081
Probabilidade de atribuição da cor azul após triagem	p5	0.00318
Taxa média de serviço de um servidor da consulta	Tsc	4
N.º de servidores nas consultas	nsc	4
Tempo final da simulação	t.end	1000

Tabela 3.5 - Parâmetros utilizados na simulação

Note-se que, sendo Ts < Tc, poderia levar à interpretação de que a fila iria ficar infinita, uma vez que o serviço seria mais lento do que a chegada de utentes. No entanto, existindo 2 servidores na fase de admissão (ns = 2) faz com que a taxa de serviço global desta fase seja de 20 utentes por hora, escoando, assim, os 15 utentes por hora que chegam ao serviço de admissão.

Como os dados fornecidos pela simulação têm um formato do tipo número e os da base de dados têm um formato do tipo hh:mm:ss foi necessário converter estes dados também para o formato numérico, para que depois fosse possível a comparação.

Para comparar as distribuições calculou-se, na simulação e na base de dados, as diferenças dos tempos de chegada à admissão inicial entre um utente e o anterior, obtendo-se, assim, o tempo entre duas chegadas consecutivas. Na base de dados, estas diferenças já tinham sido determinadas, através do cálculo de *Δha*, quando se efetuou o tratamento dos dados ainda no ficheiro do Excel, tal como está explicado na secção 3.2. A esta variável foi atribuído o nome de *Dha*, depois de converter o ficheiro para o formato .csv, conforme está também descrito na Tabela 3.2.

Por último, aplicou-se o *Teste de Kolmogorov-Smirnov*, que no software R [10] é aplicado através do comando *ks.test()* [18].

O teste devolveu um *p-value < 2.2e-16* , o que levaria a rejeitar a hipótese nula e com isso concluir que haveria evidência estatística de que as duas amostras não são provenientes da mesma distribuição. Contudo, apesar da variável tempo ser contínua na base de dados, ela está aproximada ao minuto (os primeiros registos da variável *Dha* são, 00:07:00, 00:02:00, 00:05:00, etc.) tornando-a discreta e fazendo com que o teste compare uma variável discreta com uma variável contínua, a gerada pelos dados da simulação, que, obviamente, não têm a mesma distribuição o que resulta na rejeição da hipótese nula.

Pelo facto do resultado não levar à expetativa inicial optou-se por analisar e comparar outras medidas.

Contudo, refira-se que o *Teste de Kolmogorov-Smirnov* para duas amostras analisa a distância máxima, D, entre as duas funções de distribuição empíricas. Por conseguinte, representando graficamente as duas distribuições, através da função *ecdf* (*empirical cumulative distribution function*), obtém-se o resultado da Figura 3.9, onde se pode constatar que essa distância não é significativa, levando a supor que as duas amostras podem ser provenientes da mesma população.

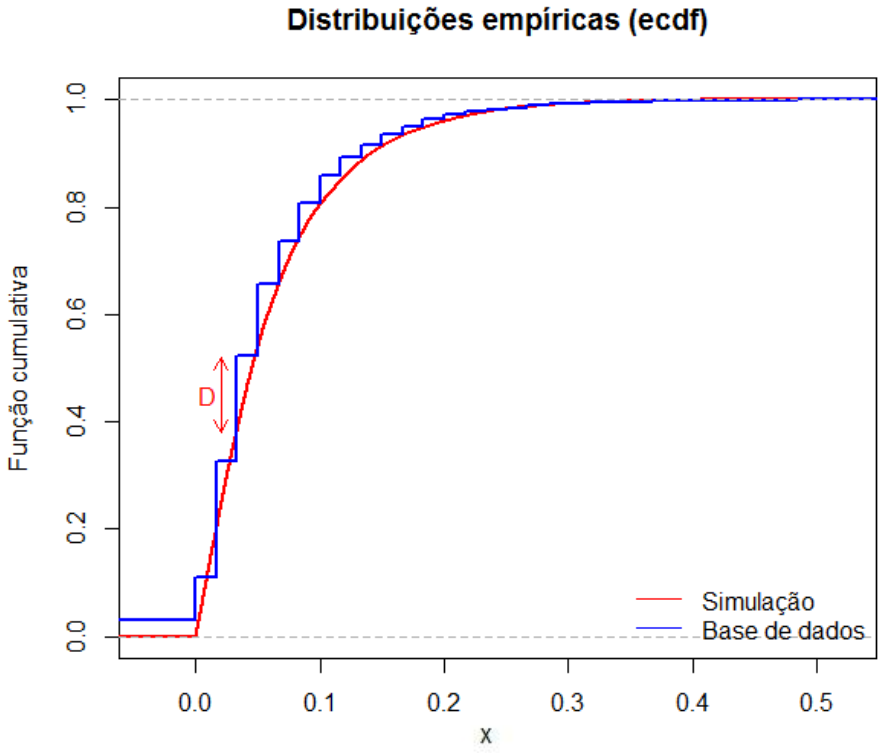


Figura 3.9 - Distribuições empíricas (coluna 2 vs. variável *Dha*)

Por outro lado, comparando algumas medidas destas duas amostras, por exemplo recorrendo às funções *summary* (resumo) e *sd* (desvio padrão) de ambas as amostras, obtiveram-se os resultados da Tabela 3.6.

```
> summary(dados_sim)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000025 0.0209700 0.0452700 0.0628500 0.0853300 0.4995000
> summary(dados_bd)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.01667 0.05000 0.06170 0.08333 0.51670
```

```
> sd(dados_sim)
[1] 0.06003213

> sd(dados_bd)
[1] 0.06195237
```

Tabela 3.6 - Algumas medidas das duas amostras (coluna 2 vs. variável *Dha*)

Analisando a tabela anterior, verificam-se resultados muito idênticos o que leva a supor que as duas amostras podem seguir a mesma distribuição.

Outra representação das duas distribuições pode ser feita através do gráfico quantil-quantil plot, que no software R é aplicado através do comando *qqplot()* [16]. Este gráfico é utilizado para analisar a distribuição dos quantis de dois conjuntos de dados. Neste gráfico, os pontos do plano são formados pelos quantis das duas amostras (dados simulados e dados reais). Se os pontos se alinharem numa reta de inclinação 1 ($x = y$), então as distribuições das duas amostras podem ser consideradas iguais. Assim, representando as duas amostras, obtém-se o gráfico da Figura 3.10, onde se pode observar que a maior parte dos pontos coincidem com a reta $x = y$ ou estão próximos dela. Face a estes resultados, mais uma vez parece não haver grande diferença entre as duas distribuições.

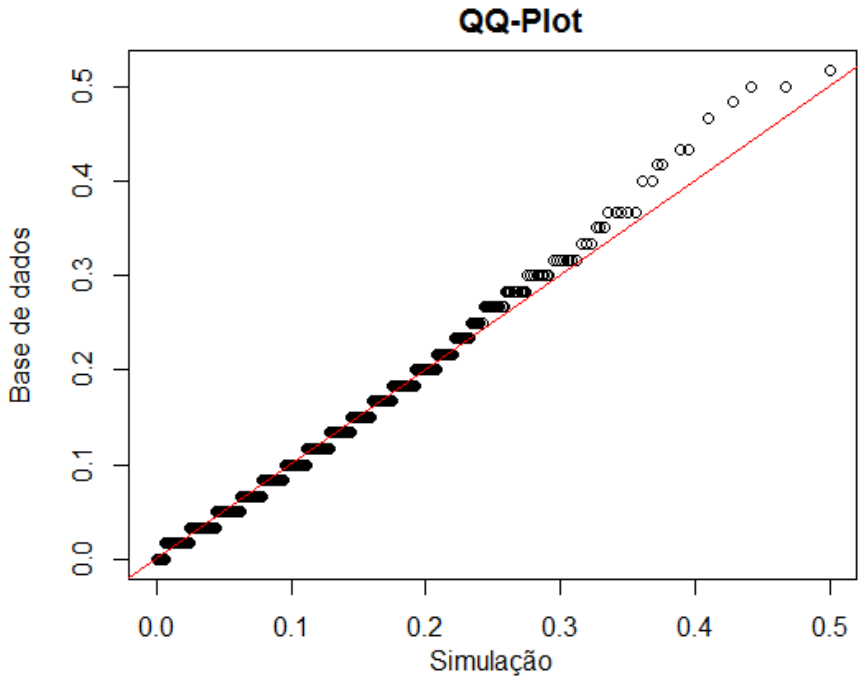


Figura 3.10 - Quantil - Quantil Plot (coluna 2 vs. variável *Dha*)

Outro teste adequado para comparar distribuições de duas amostras é o teste de ajustamento do Qui-Quadrado [21]. Este teste divide as observações em intervalos e depois compara as frequências desses intervalos. Por ser uma exigência do teste colocaram-se as duas amostras com a mesma dimensão, aplicando-se, depois, o comando *chisq.test()* [19].
O teste devolveu um *p-value = 0.4601*, não rejeitando a hipótese nula. Por isso, considerando as hipóteses iniciais, não há evidência estatística de que as duas amostras não sejam provenientes da mesma distribuição.

Concluindo, apesar do *Teste de Kolmogorov-Smirnov* rejeitar a hipótese nula, pelas razões previamente referidas, a análise gráfica, os resultados das medidas e o teste do Qui-Quadrado indicam que as duas amostras podem seguir a mesma distribuição, pelo menos de forma aproximada.

O código completo do programa da comparação destas distribuições é apresentado no Anexo II.

3.4.2 Comparação das distribuições do tempo de espera para a triagem

Tal como descrito na secção 3.2, a variável da base de dados foi calculada fazendo a diferença entre os tempos da hora de início da triagem e da hora de admissão, atribuindo-se o nome de *tet* (*tet = ht – ha)*. Neste caso, o objetivo seria comparar os tempos de espera para a triagem. No entanto, como a base de dados não tem informação sobre os tempos relativos ao fim do serviço de admissão, não foi possível fazê-lo de forma direta. Por essa razão, optou-se pela comparação da soma de dois tempos: o tempo do serviço de admissão e o tempo de espera para a triagem. Assim, de modo a que faça sentido a comparação com os dados da simulação, teve que se comparar a variável *tet*, com a diferença entre os tempos da coluna 4 (hora de início da triagem) e da coluna 2 (hora de início da admissão) da matriz de simulação. Os métodos utilizados para a comparação foram os mesmos descritos na subsecção anterior.

O *Teste de Kolmogorov-Smirnov* voltou a rejeitar a hipótese nula. Analisando o gráfico das distribuições empíricas e o gráfico quantil-quantil plot, representados na Figura 3.11,

pode-se observar que as funções de distribuição dos dois conjuntos de dados não são muito distintas e muitos dos pontos dos quantis amostrais estão próximos da reta $x = y$, embora estes se afastem um pouco da reta quando a variável assume valores mais altos.

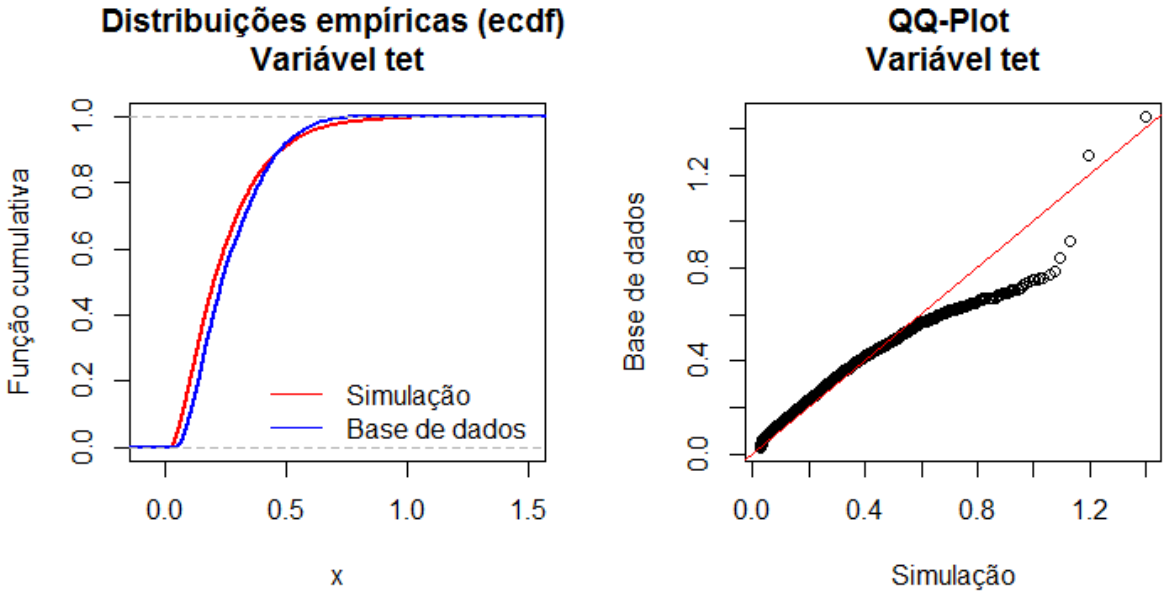


Figura 3.11 - Comparação das distribuições ((T[4]-T[2]) vs. variável *tet*)

Pelo facto da variável *tet* ter os tempos aproximados ao segundo, ela deixou de ser discreta e passou a ser mais próxima de contínua, tal como se constata na Figura 3.11, ao contrário do que acontecia com a variável *Dha*. No entanto, aparentemente, as duas amostras não seguem exatamente a mesma distribuição. Uma opção seria recorrer à estatística não paramétrica para tentar estimar a distribuição de cada um dos diferentes tempos do sistema. Contudo, uma vez que a base de dados não permite obter o tempo de serviço, pelas razões anteriormente apontadas, não foi possível usar a estatística não paramétrica de modo a aferir a distribuição de cada um dos tempos.

Face a esta impossibilidade optou-se por tentar modelar e comparar algumas medidas destas duas amostras. Isto é, manteve-se na modelação a distribuição exponencial, dado ser impossível deduzir a distribuição dos tempos, e estimou-se os valores dos parâmetros de forma a aproximar, dentro do possível, algumas medidas tais como a média, quartis, extremos e desvio padrão. Assim, recorrendo às funções *summary* e *sd*, obtiveram-se os resultados da Tabela 3.7.

```
> summary(dados_sim1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03007 0.12170 0.20340 0.24840 0.33070 1.39900
> summary(dados_bd1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0300  0.1558  0.2369  0.2702  0.3608  1.4490
```

```
> sd(dados_sim1)
[1] 0.1728463

> sd(dados_bd1)
[1] 0.1473049
```

Tabela 3.7 - Algumas medidas das duas amostras ((T[4]-T[2]) vs. variável *tet*)

Analisando a tabela anterior, verifica-se que os resultados são aproximados, levando a supor que as duas amostras podem seguir distribuições aproximadas. Refira-se que estes resultados foram obtidos após se somar 0.03 aos dados da simulação, de modo a aproximá-los dos dados da base de dados. Este tempo pode ser interpretado como um tempo que os utentes gastam enquanto circulam entre as fases do sistema.

O teste de ajustamento do Qui-Quadrado devolveu um *p-value = 0.3167*, não rejeitando a hipótese nula, não havendo, por isso, evidência estatística de que as duas amostras não sejam provenientes da mesma distribuição.

Concluindo, a análise gráfica, os resultados das medidas e o teste do Qui-Quadrado indicam que as duas amostras podem seguir a mesma distribuição, pelo menos de forma aproximada, sendo os valores encontrados para os parâmetros da simulação minimamente aceitáveis.

O código completo do programa da comparação destas distribuições é apresentado no Anexo II.

3.4.3 Comparação das distribuições do tempo de espera para a consulta

Neste caso, a análise foi feita com a diferença dos tempos entre a coluna 7 (hora de entrada na consulta) e a coluna 4 (hora de início da triagem) da matriz T de simulação e comparando-a com a variável *tec* da base de dados. A variável *tec* (*tec = ho – ht*), tal como descrito na secção 3.2, representa a diferença entre os tempos da hora de observação, ou seja, a hora de entrada na consulta, e da hora de início da triagem, coincidindo, assim, com a diferença de tempos da matriz de simulação. Tal como no tópico anterior, os métodos usados foram os descritos na subsecção 3.4.1.

As dificuldades na comparação das distribuições referidas para a variável *tet* mantiveram-se para a variável *tec*. Contudo, os valores encontrados para os parâmetros da simulação, descritos na Tabela 3.5, acabam por ser minimamente aceitáveis, usando-se, por isso, a mesma metodologia já descrita aquando da análise da variável *tet*.

Pelas mesmas razões da variável anterior, foi também somado aos dados da simulação o valor de 0.0025. Retiraram-se ainda 14 valores (correspondentes a menos de 0.1%) dos dados da simulação que estavam muito afastados dos restantes, considerando apenas os menores do que 8.2, fazendo com que os dados das duas amostras se aproximassem.

Os gráficos das distribuições empíricas e o gráfico quantil-quantil plot estão representados na Figura 3.12.

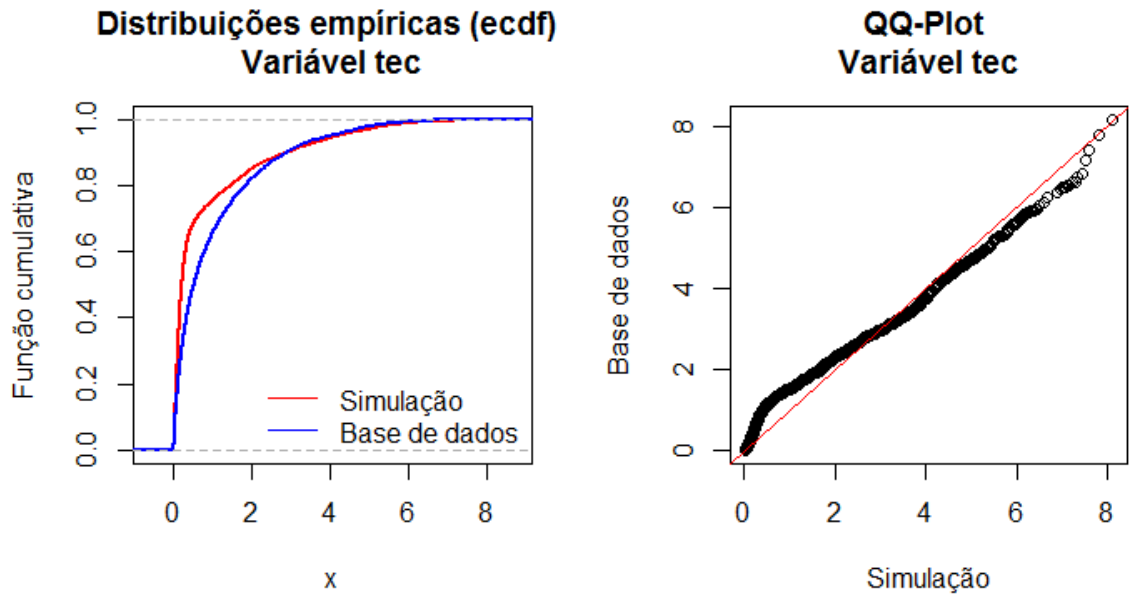


Figura 3.12 - Comparação das distribuições ((T[7]-T[4]) vs. variável *tec*)

Recorrendo às funções *summary* e *sd*, obtiveram-se os resultados da Tabela 3.8.

```
> summary(dados_sim2)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002504 0.094130 0.213300 0.854300 0.944000 8.125000
> summary(dados_bd2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0025  0.1589  0.5186  1.0510  1.4870  8.1450
```

```
> sd(dados_sim2)
[1] 1.373135

> sd(dados_bd2)
[1] 1.286707
```

Tabela 3.8 - Algumas medidas das duas amostras ((T[7]-T[4]) vs. variável *tec*)

Apesar do teste de ajustamento do Qui-Quadrado ter devolvido um *p-value = 0.2675*, não rejeitando a hipótese nula, nota-se, quer pelos gráficos, quer pelos resultados das medidas,

diferenças ligeiramente mais acentuadas em relação às variáveis anteriores. No entanto, e após diversos testes, foi com os valores dos parâmetros anteriormente referidos que se conseguiu a melhor aproximação entre a distribuição estatística da simulação e da base de dados. Todavia, saliente-se que a distribuição que caracteriza os tempos de serviço não é a exponencial, pois não é possível aproximar todos os parâmetros simultaneamente, uma vez que, quando um é tido em consideração os restantes pioram.

O código completo do programa da comparação destas distribuições é apresentado no Anexo II.

Admitindo que os valores testados para os parâmetros da simulação podem ser encarados como minimamente aceitáveis de forma a aproximar as distribuições das duas amostras, nas várias fases do sistema, efetuaram-se de seguida, várias simulações, alterando os parâmetros, de modo a testar eventuais ganhos que possam contribuir para uma melhoria da eficiência do sistema. São essas simulações e as respetivas análises e resultados que estão descritos no capítulo seguinte. Contudo, reiteramos a inexistência de dados que permitam uma análise não paramétrica através da qual se poderia estimar a função de distribuição. Deste modo, foi utlizada a usual distribuição exponencial procurando-se, unicamente, os valores dos parâmetros que melhor aproximam os dados simulados dos dados reais.

4. Análise de Resultados

Neste capítulo, pretendeu-se fazer uma análise de sensibilidade, criando cenários diversos através da alteração de vários parâmetros do modelo. Foi também testada a estabilidade do modelo e apresentados resultados e as respetivas análises. O código do programa usado para obter os resultados é apresentado no Anexo III.

Os resultados obtidos tiveram por base a criação das variáveis que se explicam em seguida.

- *tea*: média do tempo de espera, de cada utente, até iniciar o registo na admissão inicial. Obtida através do cálculo da média da diferença entre a coluna 2 (hora de início da admissão) e a coluna 1 (hora de chegada ao hospital).

- *tsa*: média do tempo que cada utente esteve na admissão, ou seja, a média do tempo de serviço da admissão. Obtida através do cálculo da média da diferença entre a coluna 3 (hora de fim da admissão) e a coluna 2 (hora de início da admissão).

- *tet*: média do tempo de espera, de cada utente, desde que saiu da admissão até iniciar a triagem. Obtida através do cálculo da média da diferença entre a coluna 4 (hora de início da triagem) e a coluna 3 (hora de fim da admissão).

- *tst*: média do tempo que cada utente esteve na triagem, ou seja, a média do tempo de serviço da triagem. Obtida através do cálculo da média da diferença entre a coluna 5 (hora de fim da triagem) e a coluna 4 (hora de início da triagem).

- *tec*: média do tempo de espera, de cada utente, desde que saiu da triagem até iniciar a consulta. Obtida através do cálculo da média da diferença entre coluna 7 (hora de entrada na consulta) e a coluna 5 (hora de fim da triagem).

- *tsc*: média do tempo que cada utente esteve na consulta, ou seja, a média do tempo de serviço da consulta. Obtida através do cálculo da média da diferença entre a coluna 8 (hora de saída da consulta) e a coluna 7 (hora de entrada na consulta).

No sistema geral em estudo, as variáveis distribuem-se como se mostra na Figura 4.1.

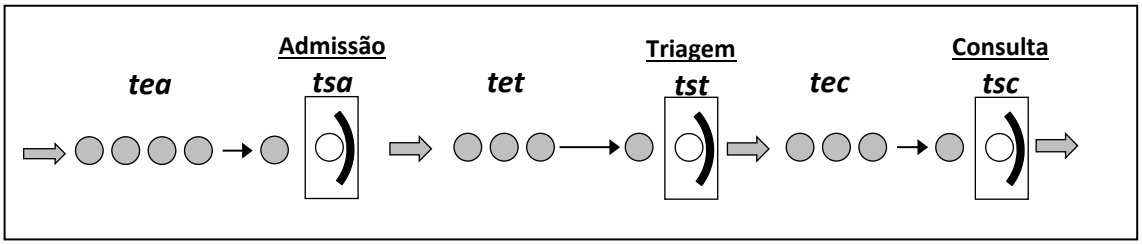


Figura 4.1 - Variáveis em análise no modelo

4.1. Alteração do número de servidores

Alterando o número de servidores na admissão (parâmetro *ns*), e mantendo os outros parâmetros nos valores definidos na Tabela 3.5, obtiveram-se os resultados da Tabela 4.1.

```
> tabela_ns
              tea         tsa         tet         tst         tec         tsc
ns=1 1.731903e+02 0.10139013 0.0443828 0.04764772 0.05135144 0.2494270
ns=2 1.161184e-01 0.09976613 0.1076443 0.04692426 0.71581846 0.2490607
ns=3 1.484662e-02 0.09970685 0.1137033 0.04780375 0.73548688 0.2504978
ns=4 2.557755e-03 0.09929135 0.1236386 0.04765309 0.94214117 0.2500997
```

Tabela 4.1 - Alterações em *ns* (número de servidores na admissão)

A alteração do parâmetro *ns*, tem, sobretudo, influência na variável *tea*, porque é nessa coluna que se notam diferenças significativas. Para *ns=1* também se verificam oscilações nas variáveis *tet* e *tec*, em relação aos outros valores de *ns*. O facto de existir apenas 1 servidor origina tempos de espera elevados, fazendo com que, depois da admissão, os utentes cheguem aos serviços seguintes com intervalos de tempos muito maiores e por isso não encontram filas de espera, daí os tempos de espera reduzidos no serviço de triagem e no de consulta. As ligeiras oscilações que as outras variáveis sofrem devem-se à geração de números aleatórios de cada vez que se efetua uma simulação, sendo que, os resultados de cada *ns* são obtidos através de simulações distintas.

Assim, verifica-se que o sistema é menos eficiente se tiver apenas 1 servidor na admissão, uma vez que o tempo de espera é muito superior em relação às outras simulações. A partir de 2 servidores o sistema começa a ser aceitável e quantos mais servidores houver menor será o tempo de espera para o serviço de admissão, como expectável, contudo a diferença é cada vez mais diminuta. Inicialmente, o sistema foi definido para 2 servidores na admissão e, considerando também os custos inerentes ao aumento de servidores, parece ser, aparentemente, um valor razoável.

Alterando o número de servidores na triagem (parâmetro *nst*), e mantendo os outros parâmetros nos valores definidos na Tabela 3.5 (análise *ceteris paribus*), obtiveram-se os resultados da Tabela 4.2.

```
> tabela_nst
            tea        tsa          tet        tst       tec       tsc
nst=1 0.1078736 0.09973827 0.1173776417 0.04755882 0.5809052 0.2468611
nst=2 0.1093408 0.10047380 0.0062664304 0.04755498 0.6290216 0.2475896
nst=3 0.1295076 0.10106611 0.0006717788 0.04732822 0.7503829 0.2512193
nst=4 0.1284899 0.09937321 0.0001077397 0.04814022 0.6725222 0.2476978
```

Tabela 4.2 - Alterações em *nst* (número de servidores na triagem)

Pelas mesmas razões apontadas para o parâmetro anterior, conclui-se que o parâmetro *nst* apenas tem influência na variável *tet*. Mais uma vez se nota que à medida que aumenta o número de servidores na triagem o tempo de espera vai diminuindo e de forma acentuada, mas as diferenças vão sendo cada vez menos significativas. O sistema, inicialmente, estava definido para 1 servidor mas, neste caso, seria vantajoso aumentar o número de servidores para 3 (ou pelo menos para 2), uma vez que há uma grande diminuição no tempo de espera, tornando o sistema muito mais eficiente. Foi com 3 servidores na triagem que se efetuaram as simulações seguintes, as do último servidor.

Alterando o número de servidores na consulta (parâmetro *nsc*), com *nst=3* e mantendo os outros parâmetros (Tabela 3.5), obtiveram-se os resultados da Tabela 4.3. Para que a análise não fique muito exaustiva optou-se por apresentar os exemplos mais relevantes.

```
> tabela_nsc
             tea        tsa          tet        tst          tec       tsc
nsc=3  0.1169663 0.10044567 0.0006632300 0.04808325 63.336419934 0.2498443
nsc=4  0.1268570 0.09968589 0.0006113701 0.04765542  0.791271697 0.2511398
nsc=6  0.1194459 0.09879659 0.0008220566 0.04756400  0.024469571 0.2501825
nsc=8  0.1169734 0.09981709 0.0007501422 0.04705180  0.002674805 0.2525874
nsc=10 0.1295956 0.10122763 0.0007413870 0.04761190  0.000218647 0.2483160
```

Tabela 4.3 - Alterações em *nsc* (número de servidores na consulta)

Pelas mesmas razões já descritas, o parâmetro *nsc* tem influência apenas na variável *tec*. Com valores inferiores a 4 servidores o sistema fica ineficiente, devido ao tempo de espera para a consulta ser demasiado elevado (para *nsc*=3 verifica-se *tec*=63.3). A partir de 6 servidores o sistema fica aceitável, mas é a partir de 8 servidores que o sistema fica eficiente, porque, depois de consultar a matriz T gerada, verifica-se que a maior parte dos tempos de entrada na consulta são iguais aos tempos de fim da triagem, significando que os utentes, praticamente, não encontram fila de espera ao passar da triagem para a consulta. Como o sistema, inicialmente, estava definido para 4 servidores, será mais vantajoso

aumentar esse número para 8. Note-se que o modelo foi simplificado uma vez que apenas se considerou postos de consulta, sem ter em conta as especialidades, uma vez que não se analisou o comportamento das filas de espera para as consultas de especialidade, embora estas existam no sistema real.

Assim, assumindo estes novos parâmetros, a que chamaremos parâmetros eficientes, calculou-se o tempo médio de permanência total dos utentes no hospital, ou seja, o tempo médio decorrido desde hora de entrada dos utentes no hospital (coluna 1 da matriz de simulação) até à hora de saída da consulta (coluna 8 da matriz de simulação) e comparou-se com o tempo de permanência considerando os parâmetros iniciais. Pelo gráfico da Figura 4.2, onde foi também introduzida uma simulação com parâmetros intermédios, verifica-se que houve uma redução significativa nesse tempo.

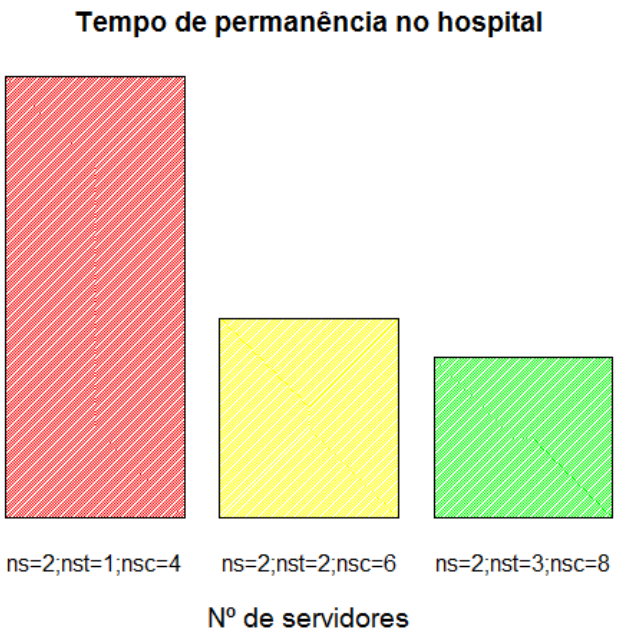


Figura 4.2 - Tempo no hospital com vários cenários

4.2. Eficiência após a triagem

Considerados os parâmetros iniciais, descritos na Tabela 3.5 e depois de calcular as médias do tempo de espera para a consulta, confirmou-se, que quanto maior for a urgência do utente, menor será o tempo de espera, como se vê pela Tabela 4.4. Recorde-se que a cor vermelha indica a maior urgência (emergente) e a cor azul a não urgente.

```
> da_triagem_para_consulta
                 Vermelho    Laranja   Amarelo    Verde     Azul
Tempo de espera 0.04090616 0.05636194 0.1223546 1.884396 8.742427
```

Tabela 4.4 - Tempos de espera depois da triagem (parâmetros iniciais)

Considerando os parâmetros eficientes, encontrados na secção 4.1 (*nst=3* e *nsc=8*), obteve-se a Tabela 4.5, onde se confirmou que os tempos de espera, depois da triagem, reduzem de forma acentuada, chegando a não haver fila de espera para os utentes classificados com cor vermelha, confirmando-se, assim, a eficiência do sistema com estes novos parâmetros.

```
> da_triagem_para_consulta
                 Vermelho     Laranja     Amarelo       Verde        Azul
Tempo de espera        0 0.001294909 0.001369415 0.002737718 0.007303118
```

Tabela 4.5 - Tempos de espera depois da triagem (parâmetros eficientes)

Em relação ao número total de utentes por cor de triagem, quer com os parâmetros iniciais, quer com os parâmetros eficientes, os valores foram idênticos, tal como se verifica na Tabela 4.6, pelo facto de estes números serem gerados aleatoriamente em função das probabilidades de atribuição de cada cor, definidas na Tabela 3.5. As pequenas diferenças que se verificaram foram devidas ao facto de se terem obtido os valores em simulações diferentes, ou seja, a eficiência do serviço não afeta os resultados da triagem.

```
> total_por_triagem
         Parâmetros iniciais Parâmetros eficientes
Vermelho                  20                    16
Laranja                  890                   871
Amarelo                 7738                  7856
Verde                   6043                  6109
Azul                      34                    37
```

Tabela 4.6 - Número de utentes por triagem

4.3. Estabilidade do modelo

Sendo a simulação um processo artificial, que retrata o comportamento de um fenómeno aleatório, esta permite observar o comportamento desse fenómeno que será tão mais fiável quantas mais vezes for repetido.

Assim, de modo a aferir a estabilidade do modelo, efetuou-se uma simulação com 100

sequências[3] e, através do comando *rbind*, do software R, complementado com o ciclo *for*, foi construída uma matriz onde cada linha corresponde aos dados gerados por cada sequência. As colunas da matriz mostram os resultados da média, da mediana e do desvio padrão, optando-se por analisar as variáveis *dados_sim* (tempo entre duas chegadas consecutivas à admissão), *dados_sim1* (diferença entre os tempos da coluna 4 e da coluna 2) e *dados_sim2* (diferença entre os tempos da coluna 7 e da coluna 4). A opção por estas três variáveis teve como finalidade medir, praticamente, todo o modelo (diferença entre chegadas, da coluna 2 à 4 e da coluna 4 à 7). Estas variáveis foram explicadas no capítulo 3 e foram introduzidas no código da comparação das distribuições, apresentado no Anexo II.

Deste modo, a matriz ficou com 100 linhas correspondentes ao número de sequências e 9 colunas, que correspondem às 3 medidas de cada uma das variáveis.

Uma parte da matriz gerada é mostrada na Tabela 4.7, onde as variáveis *dados_sim*, *dados_sim1* e *dados_sim2* foram abreviadas para *ds*, *ds1* e *ds2*, respetivamente.

```
> matriz_100seq
       Média ds Mediana ds D padrão ds Média ds1 Mediana ds1 D padrão ds1 Média ds2 Mediana ds2 D padrão ds2
 [1,] 0.06664339 0.04960049  0.06141066 0.2098018   0.1675159    0.1726988 0.8056978   0.1953066    1.4259599
 [2,] 0.06651576 0.04867305  0.06220598 0.2233958   0.1731305    0.1842743 0.8460026   0.1920506    1.8504260
 [3,] 0.06641817 0.04851561  0.06259218 0.2104113   0.1729859    0.1606571 1.2516278   0.2069185    3.2852325
 [4,] 0.06721505 0.04950226  0.06353895 0.2102965   0.1681696    0.1689521 0.6206197   0.1744618    1.2704602
 [5,] 0.06738384 0.04913519  0.06370490 0.2153588   0.1714694    0.1714329 0.7258122   0.1853363    1.5976677
 [6,] 0.06786063 0.04992106  0.06333460 0.2101297   0.1673941    0.1673035 0.5635819   0.1595673    1.0678675
 [7,] 0.06685333 0.04841046  0.06301402 0.2322282   0.1820273    0.1929397 0.8691036   0.2024221    1.7447843
 [8,] 0.06681214 0.04915225  0.06213270 0.2232335   0.1765408    0.1784097 0.6504554   0.1833713    1.4187896
 [9,] 0.06694654 0.04912872  0.06274515 0.2191716   0.1701859    0.1876012 0.8108390   0.1936147    1.4391883
[10,] 0.06668269 0.04893628  0.06247914 0.2215613   0.1743543    0.1807587 0.7018285   0.1782093    1.3839098
  ⋮         ⋮          ⋮           ⋮         ⋮           ⋮            ⋮         ⋮           ⋮            ⋮
```

Tabela 4.7 - Parte da matriz gerada através de 100 sequências

Refira-se que a mediana é obtida do conjunto de dados quando este se encontra ordenado. Contudo, a função *median*, do *software* R, já tem em conta a ordenação, não havendo a necessidade de se ordenar os dados antes de executar o comando que devolve a mediana.

Após obter a matriz das sequências, chegou-se aos resultados, através da função *summary*, representados na Tabela 4.8, que mostram um resumo de cada coluna da matriz gerada.

[3] Como curiosidade, refira-se que esta simulação demorou 17 horas até ficar concluída.

```
> summary(matriz_100seq)
    Média ds          Mediana ds        D padrão ds
 Min.   :0.06554   Min.   :0.04792   Min.   :0.06122
 1st Qu.:0.06648   1st Qu.:0.04870   1st Qu.:0.06216
 Median :0.06681   Median :0.04916   Median :0.06273
 Mean   :0.06679   Mean   :0.04912   Mean   :0.06368
 3rd Qu.:0.06716   3rd Qu.:0.04950   3rd Qu.:0.06335
 Max.   :0.06785   Max.   :0.05040   Max.   :0.08801
   Média ds1        Mediana ds1       D padrão ds1
 Min.   :0.1980   Min.   :0.1630   Min.   :0.1500
 1st Qu.:0.2136   1st Qu.:0.1710   1st Qu.:0.1689
 Median :0.2197   Median :0.1745   Median :0.1779
 Mean   :0.2193   Mean   :0.1745   Mean   :0.1775
 3rd Qu.:0.2252   3rd Qu.:0.1778   3rd Qu.:0.1852
 Max.   :0.2521   Max.   :0.1921   Max.   :0.2230
   Média ds2        Mediana ds2       D padrão ds2
 Min.   :0.5358   Min.   :0.1610   Min.   :0.9061
 1st Qu.:0.7077   1st Qu.:0.1811   1st Qu.:1.4990
 Median :0.8254   Median :0.1885   Median :1.7814
 Mean   :0.8887   Mean   :0.1911   Mean   :2.0052
 3rd Qu.:0.9992   3rd Qu.:0.2000   3rd Qu.:2.2972
 Max.   :2.1967   Max.   :0.2540   Max.   :5.8353
```

Tabela 4.8 - Resumo da matriz gerada com 100 sequências

Analisando a tabela anterior concluiu-se que, na maior parte das colunas, a diferença entre o mínimo e o máximo é pouco significativa, que a mediana é igual ou muito próxima da média e que a amplitude interquartil (diferença entre o terceiro e o primeiro quartil) é pequena. Estas características mostram que não existe grande variabilidade dos dados em cada coluna, indicando que o sistema é estável. Contudo, é também notório que a variável *ds2* apresenta uma maior variabilidade que as restantes variáveis, como se pode constatar pelas distâncias entre o máximo e as restantes medidas (nomeadamente na média e no desvio padrão).

Salienta-se, assim, a importância da estabilidade de uma simulação quando se aumenta o número de sequências, de modo a obter uma descrição mais fiável das características do sistema em estudo, destacando o papel que a simulação pode desempenhar na análise dos resultados.

5. Conclusão

Na área da saúde, as decisões sobre como e quando colocar pessoal, equipamentos, camas, e outros recursos, a fim de minimizar os atrasos sofridos pelos pacientes, são, muitas vezes, ainda mais difíceis de tomar do que em outros setores de serviços, devido a limitações de custos por um lado e consequências, potencialmente graves, de atrasos por outro lado. Portanto, é imperativo que essas decisões sejam baseadas na maior informação possível e assentes nas melhores metodologias disponíveis de forma a analisar o impacto das várias alternativas [8].

Os serviços de urgência representam um papel fulcral no Serviço Nacional de Saúde. Estes serviços são destinados a emergências e pacientes em situações urgentes, sendo normal existir uma grande afluência de pacientes no serviço e, consequentemente, originar problemas devidos a longos tempos de espera, provocando descontentamento dos pacientes.

Um serviço de urgência que se preocupe com os tempos de espera dos pacientes pode adotar medidas para os melhorar, caso estes não sejam aceitáveis. Esta ideia foi a base deste estudo e é a pensar na eficiência do serviço de urgência e no benefício dos pacientes que foi desenvolvido um modelo de simulação, através do software R, que permitiu criar vários cenários e com isso testar qual a melhor solução. A grande vantagem desta abordagem reside na profundidade de conhecimento passível de ser adquirido pela sua modelação, capaz de proporcionar um modelo de simulação preciso e próximo da realidade, concebendo resultados mais fidedignos. A simulação é uma das técnicas mais populares de operações de investigação, por ser uma ferramenta flexível, poderosa e intuitiva. Em questão de segundos ou minutos, o modelo pode simular até mesmo anos de operação de um sistema típico, gerando uma série de observações estatísticas sobre o desempenho do sistema ao longo desse período [20].

Neste estudo, não foram considerados os fatores económicos que devem estar associados, também, à eficiência de um sistema. Apenas se procurou encontrar o melhor desempenho do modelo associado ao sistema de filas de espera do estudo de caso. Considerando os custos associados ao sistema, não se pode dizer que a solução encontrada seja a solução ótima, uma vez que não se teve em conta todo o conjunto de fatores que influenciam as

tomadas de decisão. Contudo, mesmo sem os cálculos dos custos inerentes ao processo, a possível solução mais eficiente, teve em conta, intuitivamente, esse fator.

Com a simulação pretendeu-se, numa primeira fase, estimar, estatisticamente, os parâmetros pretendidos, constituindo um aspeto importante no contexto das filas de espera e, numa segunda fase, encontrar os parâmetros que melhor se adequavam ao sistema em estudo. A vantagem reside no facto de se poderem modelar, através da simulação, sistemas muito complexos, cuja análise seria difícil ou mesmo impossível de outra forma. A desvantagem consiste no facto de os resultados obtidos não fornecerem soluções exatas, mas estimativas, que variam frequentemente, embora com valores aproximados, de simulação para simulação.

Os dados operacionais necessários para os parâmetros de entrada de um modelo de filas de espera estão, muitas vezes, indisponíveis nos serviços de saúde [8]. Especificamente, neste estudo, embora os dados da hora de chegada do utente às várias fases do sistema estejam gravados, a hora de chegada ao hospital e os tempos de serviço não estão registados. Por exemplo, pelos dados fornecidos, apenas se sabe a hora de chegada dos utentes ao servidor mas não se sabe a que horas concluíram esse serviço. Este facto constituiu uma limitação para o estudo, impedindo a tradução fiel do sistema real na simulação o que, consequentemente, conduziu também a restrições na análise dos resultados.

Assim, uma análise de um sistema de filas de espera exige um esforço de recolha de dados para estimar, por exemplo, o tempo que um prestador de cuidados de saúde gasta com um paciente. Por outro lado, quanto mais comum se torna o uso de sistemas de tecnologia da informação nos cuidados de saúde, mais este tipo de dados deveria estar também facilmente disponível.

Neste estudo, ainda não tivemos disponíveis os tempos que nos permitissem modelar a distribuição do tempo de cada serviço, uma clara restrição do presente estudo. Contudo, todos os dados necessários foram inseridos no modelo de simulação, procurando analisar-se o sistema, de modo a fazê-los coincidir com os dados reais existentes, nomeadamente aquando da análise das distribuições estatísticas.

Considerando os fatores descritos, as simulações e as análises efetuadas no capítulo 4, conclui-se que os parâmetros que representam um sistema mais eficiente no serviço de urgência da unidade de saúde do HSA, são os descritos na Tabela 5.1.

Parâmetros ideais no sistema de filas de espera do HSA	
Descrição	**Valor**
Nº de servidores na admissão	2
Nº de servidores na triagem	3
Nº de servidores nas consultas	8

Tabela 5.1 - Parâmetros ideais

O modelo de simulação mostrou ser uma ferramenta adequada quando se pretende acompanhar o fluxo dos pacientes através de um serviço hospitalar e estimar os efeitos das decisões. Com este modelo de simulação é possível estimar um grande número de medidas de desempenho, relacionadas com a ocupação de recursos e com os tempos de espera dos utentes. É possível estimar, também, as alterações dessas medidas em função das decisões tomadas.

Para além dos dados, uma análise de filas de espera de uma unidade de saúde requer a identificação de outras medidas que também podem ser importantes para um serviço de excelência. Estas medidas devem refletir a perspetiva do paciente, bem como a realidade clínica [8]. Por exemplo, chegadas de utentes ao hospital com problemas não urgentes podem não necessitar de cuidados dentro de uma hora ou mais. Na perspetiva clínica será um tempo de espera razoável mas irá resultar em altos níveis de insatisfação por parte do paciente, podendo até levá-lo ao abandono do sistema, o que poderia resultar em perda de receita. Decidir sobre o que poderá ser um padrão de espera razoável numa instalação específica de cuidados de saúde não é trivial devido à falta de conhecimento de ambas as expectativas, a do paciente e a do impacto que os atrasos podem ter nos resultados clínicos para a maioria dos problemas de saúde.

Em resumo, os gestores de saúde devem estar conscientes da necessidade de usar os seus recursos de forma tão eficiente quanto possível, a fim de continuar a assegurar que as suas instituições sobrevivam e prosperem. Com este estudo tentou demonstrar-se que uma gestão eficaz do sistema de filas de espera é essencial para este objetivo, bem como para a melhoria da capacidade de atendimento, em tempo adequado, dos pacientes. No entanto, a gestão eficaz da capacidade de atendimento deve lidar com complexidades como os compromissos entre a flexibilidade e a qualidade da assistência, os tipos de pacientes, os intervalos entre chegadas variados e imprevistos, e, muitas vezes, as diferentes perspetivas dos administradores, médicos, enfermeiros e pacientes. Todos estes fatores são desafios

que afetam a capacidade dos gestores hospitalares para controlar os custos e melhorar a qualidade da prestação de cuidados de saúde. Para enfrentar esses desafios, os gestores devem estar informados, através dos dados e de medidas de desempenho do sistema, de modo a usar esses dados e medidas em modelos, para obter análises que não podem ser obtidas a partir da experiência e da intuição. A análise de sistemas de filas de espera é uma das ferramentas mais práticas e eficazes para a compreensão desses sistemas, podendo auxiliar a tomada de decisões na gestão de recursos críticos, razão pela qual deve tornar-se tão amplamente utilizada na comunidade médica como o é em outros setores de serviços.

5.1. Trabalho futuro

Face às limitações, já referidas, da base de dados do hospital, seria importante o aperfeiçoamento dos dados contidos nessa base de dados. Assim, será crucial obter dados dos tempos de cada um dos serviços, de modo a obter, através da estatística não paramétrica, a distribuição aproximada que caracteriza esses tempos. Outra informação que não consta na base de dados é a hora de chegada de cada utente ao hospital. Por isso, esse aperfeiçoamento deve passar por adotar técnicas que permitam recolher todos os dados necessários ou então, usando um método mais exaustivo, recolher esses dados através de um processo manual, um trabalho que necessitará de tempo e disponibilidade. Neste caso, o trabalho de campo iria exigir autorizações prévias e um planeamento adequado de forma a ser possível obter dados para os diferentes meses do ano, os diferentes dias da semana e as diferentes horas do dia.

Outra sugestão para um trabalho futuro será efetuar o mesmo tipo de estudo noutros serviços do hospital ou até em outras unidades hospitalares.

Por último, será interessante também aperfeiçoar o modelo de simulação, criando uma aplicação que permita a introdução de parâmetros, de forma rápida, por exemplo quando há alteração do número de médicos ou quando o número de chegadas de utentes ao hospital cresce rapidamente, de forma a observar de imediato quais as consequências no sistema.

Bibliografia

[1] Bhat, U.N. (2008). *An Introduction to Queueing Theory - Modeling and Analysis in Applications*, Statistics for Industry and Technology, Birkhäuser.

[2] Bolen,T.; Mulugeta D.; Greenfield J.; Conkey L. (2014). *An Investigation of the Kolmogorov-Smirnov Two Sample Test using SAS®*, *Cardinal Health*.

[3] Bronsom, R.; Naasimuthu, G. (2000). *Investigação Operacional* (2ª ed.), McGraw Hill.

[4] Cabral, M. (coord.); Silva, P. e Mendes, H. (2002). *Saúde e Doença em Portugal* (2ª ed.), Imprensa de Ciências Sociais, Lisboa.

[5] Carvalho, J. C.; Ramos, T. (2013). *Logística na Saúde* (2ª ed.), Edições Sílabo.

[6] Chase, R. B.; Jacobs, F. R.; Aquilano, N. J. (2004). *Operations Management for Competitive Advantage*, McGraw Hill, New York.

[7] Gross, D.; Shortle, J. F.; Thompson, J. M. & Harris, C. M. (2008). *Fundamentals of Queueing Theory* (4th ed.), John Wiley & Sons.

[8] Hall, Randolph W. (Ed.) (2006). Patient Flow: Reducing Delay in Healthcare Delivery, *International Series in Operations Research & Management Science* **91**, Springer-Verlag.

[9] Hillier, F. S.; Lieberman, G. J. (2005). *Introduction to Operations Research* (8th ed.), McGraw Hill, New York.

[10] Institute for Statistics and Mathematics (2015). *The R Project for Statistical Computing*. Disponível em: http://cran.r-project.org/ [Acedido em março 2015].

[11] Jones, O.; Maillardet, R. and Robinson, A. (2009). *Introduction to Scientific Programming and Simulation Using R*, CRC Press, Taylor & Francis Group.

[12] Justel, A.; Peña, D. and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit, *Statistics & Probability Letters*, **35**, 251-259.

[13] Kulkarni, V.G. (2011). *Introduction to Modeling and Analysis of Stochastic Systems*, Springer Texts in Statistics (2nd ed.), Springer.

[14] Little, J. D. C. (1961). A proof for the Queueing Formula: L=λW, *Operations Research,* vol. 9 (3), 383–387.

[15] Müller, D. (2007). *Processos Estocásticos e Aplicações*, Almedina.

[16] NIST/SEMATECH (2015). *e-Handbook of Statistical Methods*. Disponível em: http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm [Acedido em março 2015].

[17] Ozcan, Y.A. (2009). *Quantitative Methods in Health Care Management: Techniques and Applications*, 2nd Edition, Wiley.

[18] R Documentation, *Kolmogorov-Smirnov Tests*. Disponível em: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/ks.test.html [Acedido em março 2015].

[19] R Documentation, *Pearson's Chi-squared Test for Count Data*. Disponível em: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/chisq.test.html [Acedido em março 2015].

[20] Sharma, A. K.; Kumar, R.; Sharma, G. K. (2013). Queueing Theory Approach with Queueing Model: A Study, *International Journal of Engineering Science Invention,* vol. 2, n.º 2, 01-11.

[21] Taylor, John R. (1997). *An introduction to error analysis* (2nd ed.), Sausalito,California: University Science Books.

[22] Venables, W.N.; Smith, D.M. and the R Core Team (2015). *An Introduction to R*. Disponível em: http://cran.r-project.org/doc/manuals/R-intro.pdf.

[23] William J. Conover (1971). *Practical Nonparametric Statistics*, John Wiley & Sons, New York.

[24] Winston, W. (2003). *Operation Research: Applications and Algorithms*, (4th. Ed.), Duxbury Press.

Anexo I – Código da simulação

```
###################
# Variáveis globais
###################
t.end <- 1000     # Duração das simulações
seq = 1           # Número de réplicas (sequências)
Tc <- 15          # Taxa média de chegada dos clientes à admissão
Ts <- 10          # Taxa média de serviço de um servidor da admissão
ns <- 2           # Número de servidores na admissão
nt <- rep(0,seq)  # Número de utentes que vão entrar na fila de espera da
                  # triagem
Tst <- 21         # Taxa média de serviço de um servidor da triagem
nst <- 1          # Número de servidores na triagem
Tsc <- 4          # Taxa média de serviço de um servidor da consulta
nsc <- 4          # Número de servidores na consulta

# Probabilidades da triagem de Manchester
p1=0.00145; p2=0.05984; p3=0.52472; p4=0.41081; p5=0.00318

triagem <- function(){a=runif(1)
            if (a<p1) {return(1)}              # Vermelho
            else if (a<p1+p2) {return(2)}      # Laranja
            else if (a<p1+p2+p3) {return(3)}   # Amarelo
            else if (a<p1+p2+p3+p4) {return(4)}# Verde
            else {return(5)}                   # Azul
            }

##########
# Matriz T
##########
# Tem 8 colunas que caracterizam cada utente:
# C1 - Hora de chegada ao hospital
# C2 - Hora de início da admissão
# C3 - Hora de fim da admissão
# C4 - Hora de início da triagem
# C5 - Hora de fim da triagem
# C6 - Resultado da triagem
# C7 - Hora de entrada na consulta
# C8 - Hora de saída da consulta

T=rbind()              # Construção da matriz T

####################
#Sistema de admissão
####################
nu <- 0                # Número de utentes que já chegaram ao hospital
t.clock <- 0           # Relógio
n <- 0                 # Número de clientes no sistema de admissão
t1 = rexp(1, Tc)       # Tempo da primeira chegada à admissão
```

```
t2 = t.end              # Inicialização do t2
while (min(t1,t2)< t.end)
{     if (t1<t2)        # Chegada de novo utente à fila da admissão
      {
            nu=nu+1
            T=rbind(T,c(t1,0,t.end,0,t.end,0,t.end,t.end))
            n=n+1
            t.clock=t1
            t1= t1 + rexp(1, Tc)    # Próximo cliente
            if (n<=ns) {T[nu,2] = t.clock
                  T[nu,3] = t.clock + rexp(1, Ts) #verifica se
# existe servidor livre
                  T=rbind(T[order(T[,3]),])
                  t2=T[nu-n+1,3]}
      } else      # Saída de um utente com a admissão realizada
      {
            n=n-1
            t.clock=t2
            T=T[order(T[,1]),]
            if (n>ns-1) {T[nu-n+ns,2] = t.clock
                  T[nu-n+ns,3]= t.clock + rexp(1, Ts)
                  T=T[order(T[,3]),]}
            if (n>0) {t2=T[nu-n+1,3]} else {t2=t.end}
      }
}
T=T[1:(nu-n),]    # Retira os valores que não entraram na admissão até
                  # t.end

####################
#Sistema de triagem
####################
t.clock <- 0      # Relógio
n <- 0            # Número de clientes no sistema de triagem
nut <- 0;         # Número de utentes que chegaram à triagem
t1 = T[1,3]       # Tempo da primeira chegada à triagem
t2 = t.end        # Inicialização do t2
dimensao<-nrow(T) # Número de linhas
while (min(t1,t2)< t.end)
{     if (t1<t2)  # Chegada de novo utente à fila da triagem
      {
            nut=nut+1
            n=n+1
            t.clock=t1
            T=T[order(T[,3]),]
            if (nut<dimensao) {t1= T[nut+1,3]} else {t1=t.end}
            if (n<=nst) {T[nut,4] = t.clock
                  T[nut,5] = t.clock + rexp(1, Tst)
                  T[nut,6] = triagem()
                  T=T[order(T[,5]),]
                  t2=T[nut-n+1,5]}
      } else # Saída de um utente com a triagem realizada
      {
            n=n-1
            t.clock=t2
```

```
            T=T[order(T[,3]),]
            if (n>nst-1) {T[nut-n+nst,4] = t.clock
                  T[nut-n+nst,5]= t.clock + rexp(1, Tst)
                  T[nut-n+nst,6] = triagem()
                  T=T[order(T[,5]),]}
            if (n>0) {t2=T[nut-n+1,5]} else {t2=t.end}
      }
}
T=T[1:(nut-n),]   # Retira os valores que não entraram na triagem até
                  # t.end

####################
#Sistema de Consulta
####################
t.clock <- 0      # Relógio
n <- 0            # Número de clientes no sistema de consulta
nuc <- 0;         # Número de utentes que chegaram à consulta
t1 = T[1,5]       # Tempo da primeira chegada à consulta
t2 = t.end        # Inicialização do t2
dimensao<-nrow(T) # Número de linhas
while (min(t1,t2)< t.end)
{     if (t1<t2)  # Chegada de novo utente à fila da consulta
      {
            nuc=nuc+1
            n=n+1
            t.clock=t1
            T=T[order(T[,7],T[,5]),]
            if (nuc<dimensao) {t1= T[nuc+1,5]} else {t1=t.end}
            if (n<=nsc) {T[nuc,7] = t.clock
                  T[nuc,8] = t.clock + rexp(1, Tsc)
                  T=T[order(T[,8]),]
                  t2=T[nuc-n+1,8]}
      } else # Saída de um utente com a consulta realizada
      {
            n=n-1
            t.clock=t2
            T=T[order(T[,7],T[,5]),]
            if (n>nsc-1) {
                  aux=order(T[(nuc-n+nsc):nuc,6],T[(nuc- n + nsc):
                                                   nuc,5])[1]
                  T[nuc-n+nsc+aux-1,7] = t.clock
                  T[nuc-n+nsc+aux-1,8]= t.clock + rexp(1, Tsc)
                  T=T[order(T[,8]),]}
            if (n>0) {t2=T[nuc-n+1,8]} else {t2=t.end}
      }
}

T=T[250:(nuc-n),]      # Retira valores menos estáveis da matriz

T=T[order(T[,1]),]     # Ordena pela hora de chegada ao hospital
```

Anexo II – Código da comparação das distribuições

```
# ------------------------------------------------------------------------
# Comparação da coluna 2 da matriz de simulação com a variável Dha do
# subconjunto da base de dados
# ------------------------------------------------------------------------

###########
# Simulação
###########

dados_sim<-diff(T[,2])  # Diferença entre linhas da coluna 2
                        # (registo seguinte - registo anterior)

dados_sim<-dados_sim[dados_sim<0.5] # retira 1 valor muito afastado dos
                                    # restantes

###############
# Base de dados
###############

h<-read.table('bd parcial.csv',header=TRUE,sep=';') # Carrega subconjunto
                                                    # da base de dados
Dha_char<-format(h$Dha, format = "%H:%M:%S")   # Converte para H:M:S com
                                               # formato caracter

Dha_num_dec<-sapply(strsplit(Dha_char,":"), # Aplica a função a cada
# elemento da lista e separa os elementos de caracteres pelos ":"
     function(x){
     x<-as.numeric(x)
     x[1]+x[2]/60}) # Converte para numérico. Coloca a posição 1(hora) e
                    # divide a posição 2 por 60 minutos para dar decimal

dados_bd<-Dha_num_dec

# Teste Kolmogorov-Smirnov
dados_sim<-sort(dados_sim)   # Ordenar os dados de ambas as amostras por
dados_bd<-sort(dados_bd)     # ordem crescente
ks.test(dados_sim,dados_bd)

# Teste do Qui-Quadrado
dados_sim_qq<-head(dados_sim,length(dados_bd))       # Tamanhos iguais
chisq.test(dados_sim_qq,dados_bd)

##########
# Gráficos
##########

# Distribuições empíricas

plot(ecdf(dados_sim), do.points=FALSE, verticals=TRUE, pch=20, cex=1.25,
```

```
lwd=2,col="red",ylab="Função cumulativa", main="Distribuições empíricas
(ecdf)")
lines(ecdf(dados_bd), do.points=FALSE, verticals=TRUE, lwd =2,
col="blue")
legend("bottomright", c("Simulação", "Base de dados"), lty = 1,
col=c("red", "blue"), bty = "n")
arrows(0.02,0.38,0.02,0.52,length=0.1,col="red",code=3)
text(0.01,0.45,labels="D",col="red")

# QQ-Plot

qqplot(dados_sim,dados_bd,xlab="Simulação",ylab="Base de dados",
main="QQ-Plot")
abline(0,1,col="red")

########
# Resumo
########

summary(dados_sim)
summary(dados_bd)
sd(dados_sim)
sd(dados_bd)

# ---------------------------------------------------------------------
# Comparação da diferença entre a coluna 4 e a coluna 2 da matriz de
# simulação com a variável tet do subconjunto da base de dados
# ---------------------------------------------------------------------

############
# Simulação
############

dados_sim1<-T[,4]-T[,2]      # Diferença entre coluna 4 e coluna 2
dados_sim1<-dados_sim1+0.03  # Aproximar ao mínimo da base de dados

###############
# Base de dados
###############

tet_char<-format(h$tet, format = "%H:%M:%S")   # Converte para H:M:S com
                                               # formato caracter

tet_num_dec<-sapply(strsplit(tet_char,":"),  # Aplica a função a cada
# elemento da lista e separa os elementos de caracteres pelos ":"
      function(x){
      x<-as.numeric(x)
      x[1]+x[2]/60+x[3]/3600}) # Converte para numérico. Coloca a
# posição 1 (hora), divide a posição 2 por 60 minutos e a posição 3  por
# 3600 segundos para dar decimal

dados_bd1<-tet_num_dec
```

```
# Teste Kolmogorov-Smirnov
dados_sim1<-sort(dados_sim1) # Ordenar os dados de ambas as amostras por
dados_bd1<-sort(dados_bd1)   # ordem crescente
ks.test(dados_sim1,dados_bd1)

# Teste do Qui-Quadrado
dados_sim_qq1<-head(dados_sim1,length(dados_bd1))    # Tamanhos iguais
chisq.test(dados_sim_qq1,dados_bd1)

##########
# Gráficos
##########

par(mfrow=c(1,2))

# Distribuições empíricas e QQ-Plot

plot(ecdf(dados_sim1), do.points=FALSE, verticals=TRUE, pch=20, ex=1.25,
lwd=2, col="red",ylab="Função cumulativa", main="Distribuições empíricas
(ecdf) \nVariável tet")
lines(ecdf(dados_bd1), do.points=FALSE, verticals=TRUE, lwd =2,
col="blue")
legend("bottomright", c("Simulação", "Base de dados"), lty = 1,
col=c("red", "blue"), bty = "n")

qqplot(dados_sim1,dados_bd1,xlab="Simulação",ylab="Base de dados",
main="QQ-Plot \nVariável tet")
abline(0,1,col="red")

########
# Resumo
########

summary(dados_sim1)
summary(dados_bd1)
sd(dados_sim1)
sd(dados_bd1)

# ------------------------------------------------------------------------
# Comparação da diferença entre a coluna 7 e a coluna 4 da matriz de
# simulação com a variável tec do subconjunto da base de dados
# ------------------------------------------------------------------------

###########
# Simulação
###########

dados_sim2<-T[,7]-T[,4]       # Diferença entre coluna 7 e coluna 4
dados_sim2<-dados_sim2+0.0025 # Aproximar ao mínimo da base de dados
dados_sim2<-dados_sim2[dados_sim2<8.2]   # Eliminar valores afastados
```

```
###############
# Base de dados
###############

tec_char<-format(h$tec, format = "%H:%M:%S")   # Converte para H:M:S com
                                               # formato caracter

tec_num_dec<-sapply(strsplit(tec_char,":"),    # Aplica a função a cada
# elemento da lista e separa os elementos de caracteres pelos ":"
      function(x){
      x<-as.numeric(x)
      x[1]+x[2]/60+x[3]/3600}) # Converte para numérico. Coloca a
# posição 1 (hora), divide a posição 2 por 60 minutos e a posição 3  por
# 3600 segundos para dar decimal

dados_bd2<-tec_num_dec

# Teste Kolmogorov-Smirnov
dados_sim2<-sort(dados_sim2) # Ordenar os dados de ambas as amostras por
dados_bd2<-sort(dados_bd2)   # ordem crescente
ks.test(dados_sim2,dados_bd2)

# Teste do Qui-Quadrado
dados_sim_qq2<-head(dados_sim2,length(dados_bd2))    # Tamanhos iguais
chisq.test(dados_sim_qq2,dados_bd2)

##########
# Gráficos
##########

par(mfrow=c(1,2))

# Distribuições empíricas e QQ-Plot

plot(ecdf(dados_sim2), do.points=FALSE, verticals=TRUE, pch=20,cex=1.25,
lwd=2, col="red",ylab="Função cumulativa", main="Distribuições empíricas
(ecdf) \nVariável tec")
lines(ecdf(dados_bd2), do.points=FALSE, verticals=TRUE, lwd =2,
col="blue")
legend("bottomright", c("Simulação", "Base de dados"), lty = 1,
col=c("red", "blue"), bty = "n")

qqplot(dados_sim2,dados_bd2,xlab="Simulação",ylab="Base de dados",
main="QQ-Plot \nVariável tec")
abline(0,1,col="red")

########
# Resumo
########

summary(dados_sim2)
summary(dados_bd2)
sd(dados_sim2)
sd(dados_bd2)
```

Anexo III – Código dos resultados

```
# ----------------------------------------------------------------------
# Alterações no número de servidores da admissão (parâmetro ns)
# ----------------------------------------------------------------------

# ns = 1
tea1<-mean(T[,2]-T[,1]) # Média do tempo de espera até iniciar admissão.
tsa1<-mean(T[,3]-T[,2]) # Média do tempo de serviço da admissão.
tet1<-mean(T[,4]-T[,3]) # Média do tempo de espera da admissão à triagem
tst1<-mean(T[,5]-T[,4]) # Média do tempo de serviço da triagem.
tec1<-mean(T[,7]-T[,5]) # Média do tempo de espera da triagem à consulta
tsc1<-mean(T[,8]-T[,7]) # Média do tempo de serviço da consulta.

ns1<-rbind(tea1,tsa1,tet1,tst1,tec1,tsc1) # Construção da matriz

# ns = 2
tea2<-mean(T[,2]-T[,1])
tsa2<-mean(T[,3]-T[,2])
tet2<-mean(T[,4]-T[,3])
tst2<-mean(T[,5]-T[,4])
tec2<-mean(T[,7]-T[,5])
tsc2<-mean(T[,8]-T[,7])

ns2<-rbind(tea2,tsa2,tet2,tst2,tec2,tsc2)

# ns = 3
tea3<-mean(T[,2]-T[,1])
tsa3<-mean(T[,3]-T[,2])
tet3<-mean(T[,4]-T[,3])
tst3<-mean(T[,5]-T[,4])
tec3<-mean(T[,7]-T[,5])
tsc3<-mean(T[,8]-T[,7])

ns3<-rbind(tea3,tsa3,tet3,tst3,tec3,tsc3)

# ns = 4
tea4<-mean(T[,2]-T[,1])
tsa4<-mean(T[,3]-T[,2])
tet4<-mean(T[,4]-T[,3])
tst4<-mean(T[,5]-T[,4])
tec4<-mean(T[,7]-T[,5])
tsc4<-mean(T[,8]-T[,7])

ns4<-rbind(tea4,tsa4,tet4,tst4,tec4,tsc4)

# Construção da tabela
tabela_ns<-matrix(c(ns1,ns2,ns3,ns4),ncol=6,byrow=TRUE)
rownames(tabela_ns)<-c("ns=1","ns=2","ns=3","ns=4")
colnames(tabela_ns)<-c("tea","tsa","tet","tst","tec","tsc")
tabela_ns
```

```
# ------------------------------------------------------------------------
# Alterações no número de servidores da triagem (parâmetro nst)
# ------------------------------------------------------------------------

# nst = 1
tea1<-mean(T[,2]-T[,1])
tsa1<-mean(T[,3]-T[,2])
tet1<-mean(T[,4]-T[,3])
tst1<-mean(T[,5]-T[,4])
tec1<-mean(T[,7]-T[,5])
tsc1<-mean(T[,8]-T[,7])

nst1<-rbind(tea1,tsa1,tet1,tst1,tec1,tsc1)

# nst = 2
tea2<-mean(T[,2]-T[,1])
tsa2<-mean(T[,3]-T[,2])
tet2<-mean(T[,4]-T[,3])
tst2<-mean(T[,5]-T[,4])
tec2<-mean(T[,7]-T[,5])
tsc2<-mean(T[,8]-T[,7])

nst2<-rbind(tea2,tsa2,tet2,tst2,tec2,tsc2)

# nst = 3
tea3<-mean(T[,2]-T[,1])
tsa3<-mean(T[,3]-T[,2])
tet3<-mean(T[,4]-T[,3])
tst3<-mean(T[,5]-T[,4])
tec3<-mean(T[,7]-T[,5])
tsc3<-mean(T[,8]-T[,7])

nst3<-rbind(tea3,tsa3,tet3,tst3,tec3,tsc3)

# nst = 4
tea4<-mean(T[,2]-T[,1])
tsa4<-mean(T[,3]-T[,2])
tet4<-mean(T[,4]-T[,3])
tst4<-mean(T[,5]-T[,4])
tec4<-mean(T[,7]-T[,5])
tsc4<-mean(T[,8]-T[,7])

nst4<-rbind(tea4,tsa4,tet4,tst4,tec4,tsc4)

# Construção da tabela
tabela_nst<-matrix(c(nst1,nst2,nst3,nst4),ncol=6,byrow=TRUE)
rownames(tabela_nst)<-c("nst=1","nst=2","nst=3","nst=4")
colnames(tabela_nst)<-c("tea","tsa","tet","tst","tec","tsc")
tabela_nst
```

```
# ------------------------------------------------------------------------
# Alterações no número de servidores da consulta (parâmetro nsc)
# ------------------------------------------------------------------------

# nsc = 3
tea3<-mean(T[,2]-T[,1])
tsa3<-mean(T[,3]-T[,2])
tet3<-mean(T[,4]-T[,3])
tst3<-mean(T[,5]-T[,4])
tec3<-mean(T[,7]-T[,5])
tsc3<-mean(T[,8]-T[,7])
nsc3<-rbind(tea3,tsa3,tet3,tst3,tec3,tsc3)

# nsc = 4
tea4<-mean(T[,2]-T[,1])
tsa4<-mean(T[,3]-T[,2])
tet4<-mean(T[,4]-T[,3])
tst4<-mean(T[,5]-T[,4])
tec4<-mean(T[,7]-T[,5])
tsc4<-mean(T[,8]-T[,7])
nsc4<-rbind(tea4,tsa4,tet4,tst4,tec4,tsc4)

# nsc = 6
tea6<-mean(T[,2]-T[,1])
tsa6<-mean(T[,3]-T[,2])
tet6<-mean(T[,4]-T[,3])
tst6<-mean(T[,5]-T[,4])
tec6<-mean(T[,7]-T[,5])
tsc6<-mean(T[,8]-T[,7])
nsc6<-rbind(tea6,tsa6,tet6,tst6,tec6,tsc6)

# nsc = 8
tea8<-mean(T[,2]-T[,1])
tsa8<-mean(T[,3]-T[,2])
tet8<-mean(T[,4]-T[,3])
tst8<-mean(T[,5]-T[,4])
tec8<-mean(T[,7]-T[,5])
tsc8<-mean(T[,8]-T[,7])
nsc8<-rbind(tea8,tsa8,tet8,tst8,tec8,tsc8)

# nsc = 10
tea10<-mean(T[,2]-T[,1])
tsa10<-mean(T[,3]-T[,2])
tet10<-mean(T[,4]-T[,3])
tst10<-mean(T[,5]-T[,4])
tec10<-mean(T[,7]-T[,5])
tsc10<-mean(T[,8]-T[,7])
nsc10<-rbind(tea10,tsa10,tet10,tst10,tec10,tsc10)

# Construção da tabela
tabela_nsc<-matrix(c(nsc3,nsc4,nsc6,nsc8,nsc10),ncol=6,byrow=TRUE)
rownames(tabela_nsc)<-c("nsc=3","nsc=4","nsc=6","nsc=8","nsc=10")
colnames(tabela_nsc)<-c("tea","tsa","tet","tst","tec","tsc")
tabela_nsc
```

```
# ------------------------------------------------------------------------
# Gráfico que compara o tempo médio de permanência no hospital em 3
# simulações com parâmetros iniciais, intermédios e eficientes
# ------------------------------------------------------------------------

p_iniciais<-mean(T[,8]-T[,1])      # Média do tempo
p_intermed<-mean(T[,8]-T[,1])      # calculada em
p_eficientes<-mean(T[,8]-T[,1])    # 3 simulações distintas

tempos<-c(p_iniciais,p_intermed,p_eficientes)
names(tempos)<-c("ns=2;nst=1;nsc=4","ns=2;nst=2;nsc=6","ns=2;nst=3;
nsc=8")
barplot(tempos,col=c("red","yellow","green"),density=60,angle=45,
axes=FALSE,cex.lab=1.2,main="Tempo de permanência no hospital", xlab="Nº
de servidores")

# ------------------------------------------------------------------------
# Eficiência após a triagem
# ------------------------------------------------------------------------
# Cálculo da média do tempo de espera para a consulta após a triagem

tr1<-T[T[,6]==1,]                  # Triagem de cor vermelha
triagem1<-mean(tr1[,7]-tr1[,5])

tr2<-T[T[,6]==2,]                  # Triagem de cor laranja
triagem2<-mean(tr2[,7]-tr2[,5])

tr3<-T[T[,6]==3,]                  # Triagem de cor amarela
triagem3<-mean(tr3[,7]-tr3[,5])

tr4<-T[T[,6]==4,]                  # Triagem de cor verde
triagem4<-mean(tr4[,7]-tr4[,5])

tr5<-T[T[,6]==5,]                  # Triagem de cor azul
triagem5<-mean(tr5[,7]-tr5[,5])

# Construção da tabela

da_triagem_para_consulta<-matrix(c(triagem1,triagem2,triagem3,triagem4,
triagem5),ncol=5,byrow=TRUE)
rownames(da_triagem_para_consulta)<-c("Tempo de espera")
colnames(da_triagem_para_consulta)<-c("Vermelho","Laranja","Amarelo",
"Verde","Azul")
da_triagem_para_consulta

# Total por triagem

p_iniciais_tr<-table(T[,6])  # Totais das cores da triagem
p_eficientes_tr<-table(T[,6])# obtidos em simulações diferentes

total_por_triagem<-matrix(c(p_iniciais_tr,p_eficientes_tr),nrow=5,
ncol=2, dimnames=list(c("Vermelho","Laranja","Amarelo","Verde","Azul"),
c("Parâmetros iniciais","Parâmetros eficientes")))
total_por_triagem
```

```
# ------------------------------------------------------------------------
# Estabilidade do modelo
# ------------------------------------------------------------------------

t.end <- 1000     # Duração das simulações
seq=100           # Número de réplicas (sequências)
Tc <- 15          # Taxa média de chegada dos clientes à admissão
Ts <- 10          # Taxa média de serviço de um servidor da admissão
ns <- 2           # Número de servidores na admissão
nt <- rep(0,seq)  # Número de utentes que vão entrar na fila de espera da
                  # triagem
Tst <- 21         # Taxa média de serviço de um servidor da triagem
nst <- 1          # Número de servidores na triagem
Tsc <- 4          # Taxa média de serviço de um servidor da consulta
nsc <- 4          # Número de servidores na consulta

# Probabilidades da triagem de Manchester
p1=0.00145; p2=0.05984; p3=0.52472; p4=0.41081; p5=0.00318

triagem <- function(){a=runif(1)
            if (a<p1) {return(1)}              # Vermelho
            else if (a<p1+p2) {return(2)}      # Laranja
            else if (a<p1+p2+p3) {return(3)}   # Amarelo
            else if (a<p1+p2+p3+p4) {return(4)}# Verde
            else {return(5)}                   # Azul
            }

#########
# Matriz
#########

matriz_100seq=rbind()  # Construção da matriz das sequências

for (i in 1:seq)       # Início do ciclo de sequências de simulação
{
      T=rbind()        # Construção da matriz T

      ####################
      #Sistema de admissão
      ####################
      nu=0             # Número de utentes que já chegaram ao hospital
      t.clock <- 0     # Relógio
      n <-0            # Número de clientes no sistema de admissão
      t1= rexp(1, Tc)  # Tempo da primeira chegada à admissão
      t2=t.end         # Inicialização do t2
      while (min(t1,t2)< t.end)
      {     if (t1<t2)  # Chegada de um novo utente à fila da admissão
            {
                  nu=nu+1
                  T=rbind(T,c(t1,0,t.end,0,t.end,0,t.end,t.end))
                  n=n+1
                  t.clock=t1
                  t1= t1 + rexp(1, Tc)    # Próximo cliente
                  if (n<=ns) {T[nu,2] = t.clock
```

```
                        T[nu,3] = t.clock + rexp(1, Ts) #verifica se
#existe servidor livre
                        T=rbind(T[order(T[,3]),])
                        t2=T[nu-n+1,3]}
            } else      # Saída de um utente com a admissão realizada
            {
                  n=n-1
                  t.clock=t2
                  T=T[order(T[,1]),]
                  if (n>ns-1) {T[nu-n+ns,2] = t.clock
                        T[nu-n+ns,3]= t.clock + rexp(1, Ts)
                        T=T[order(T[,3]),]}
                  if (n>0) {t2=T[nu-n+1,3]} else {t2=t.end}
            }
      }
      T=T[1:(nu-n),]    # Retira os valores que não entraram na
                        # admissão até t.end

      ####################
      #Sistema de triagem
      ####################
      t.clock <- 0      # Relógio
      n <-0             # Número de clientes no sistema do triagem
      nut = 0;          # Número de utentes que chegaram à triagem
      t1= T[1,3]        # Tempo da primeira chegada à triagem
      t2=t.end          # Inicialização do t2
      dimensao<-nrow(T) # Número de linhas
      while (min(t1,t2)< t.end)
      {     if (t1<t2)  # Chegada de um novo utente à fila da triagem
            {
                  nut=nut+1
                  n=n+1
                  t.clock=t1
                  T=T[order(T[,3]),]
                  if (nut<dimensao) {t1= T[nut+1,3]} else {t1=t.end}
                  if (n<=nst) {T[nut,4] = t.clock
                        T[nut,5] = t.clock + rexp(1, Tst)
                        T[nut,6] = triagem()
                        T=T[order(T[,5]),]
                        t2=T[nut-n+1,5]}
            } else # Saída de um utente com triagem realizada
            {
                  n=n-1
                  t.clock=t2
                  T=T[order(T[,3]),]
                  if (n>nst-1) {T[nut-n+nst,4] = t.clock
                        T[nut-n+nst,5]= t.clock + rexp(1, Tst)
                        T[nut-n+nst,6] = triagem()
                        T=T[order(T[,5]),]}
                  if (n>0) {t2=T[nut-n+1,5]} else {t2=t.end}
            }
      }
      T=T[1:(nut-n),]         # Retira os valores que não entraram na
                              # triagem até t.end
```

```
      ####################
      #Sistema de Consulta
      ####################
      t.clock <- 0      # Relógio
      n <-0             # Número de clientes no sistema de consulta
      nuc = 0;          # Número de utentes que chegaram à consulta
      t1= T[1,5]        # Tempo da primeira chegada à consulta
      t2=t.end          # Inicialização do t2
      dimensao<-nrow(T) # Número de linhas
      while (min(t1,t2)< t.end)
      {     if (t1<t2)  # Chegada de um novo utente à fila da consulta
            {
                  nuc=nuc+1
                  n=n+1
                  t.clock=t1
                  T=T[order(T[,7],T[,5]),]
                  if (nuc<dimensao) {t1= T[nuc+1,5]} else {t1=t.end}
                  if (n<=nsc) {T[nuc,7] = t.clock
                        T[nuc,8] = t.clock + rexp(1, Tsc)
                        T=T[order(T[,8]),]
                        t2=T[nuc-n+1,8]}
            } else # Saída de um utente com a consulta realizada
            {
                  n=n-1
                  t.clock=t2
                  T=T[order(T[,7],T[,5]),]
                  if (n>nsc-1) {
                        aux=order(T[(nuc-n+nsc):nuc,6],T[(nuc-n + nsc):
                                                      nuc,5])[1]
                        T[nuc-n+nsc+aux-1,7] = t.clock
                        T[nuc-n+nsc+aux-1,8]= t.clock + rexp(1, Tsc)
                        T=T[order(T[,8]),]}
                  if (n>0) {t2=T[nuc-n+1,8]} else {t2=t.end}
            }
      }
T=T[250:(nuc-n),]            # Retira valores menos estáveis da matriz T

ds_100seq<-diff(sort(T[,2])) # Diferença entre linhas da coluna 2
ds1_100seq<-T[,4]-T[,2]      # Diferença entre coluna 4 e coluna 2
ds2_100seq<-T[,7]-T[,4]      # Diferença entre coluna 7 e coluna 4

# Construção da matriz com cálculo da média, mediana e desvio padrão

matriz_100seq=rbind(matriz_100seq,c(mean(ds_100seq),median(ds_100seq),sd
(ds_100seq),mean(ds1_100seq),median(ds1_100seq),sd(ds1_100seq),mean(ds2_
100seq),median(ds2_100seq),sd(ds2_100seq)))
}

colnames(matriz_100seq)<-c("Média ds","Mediana ds","D padrão ds", "Média
ds1","Mediana ds1","D padrão ds1","Média ds2","Mediana ds2","D padrão
ds2")

matriz_100seq
summary(matriz_100seq)       # Resumo
```