Universidade do Minho
Escola de Ciências

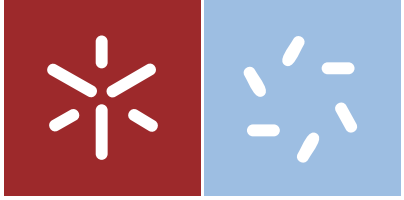
Bernardo Domingos Manuel

Filas de Espera e Aplicações

Bernardo Domingos Manuel Filas de Espera e Aplicações

UMinho | 2019

outubro de 2019



Universidade do Minho
Escola de Ciências

Bernardo Domingos Manuel

Filas de Espera e Aplicações

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Maria Conceição Soares Serra

outubro de 2019

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho

Atribuição
CC BY

https://creativecommons.org/licenses/by/4.0/

Agradecimentos

Primeiramente agradeço a Deus por ter me dado forças para terminar este trabalho.

Igualmente agradeço a minha orientadora Professora Doutora Maria da Conceição Soares Serra pela dedicação e paciência que teve na orientação deste trabalho.

Agradeço a minha esposa Yola Manuel, os meus filhos, Tchissola Manuel, Jonatã Manuel e Bernardo Manuel, pois foi com o vosso apoio, dedicação e coragem, mesmo distante deram-me coragem para terminar este trabalho.

Agradeço a minha mãe Branca Francisco Diogo e a todos os meus irmãos pela força que me deram durante a minha formação.

Tenho também de agradecer a todos os docentes do departamento de matemática e aplicações pelos conhecimentos transmitido e a todos os meus colegas do mestrado em estatística pela amizade, troca de experiências que tivemos ao longo de todo curso.

E por ultimo agradecer a todas as pessoas que direta ou indiretamente ajudaram para que este trabalho fosse uma realidade.

Declaração de integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração. Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

A teoria de filas de espera é um ramo da probabilidade aplicada e da investigação operacional. Ela encarrega-se do estudo de diferentes modelos probabilísticos que permitem descrever a evolução de uma fila de espera. Os aspetos de interesse de um modelo de fila de espera são, por exemplo: o processo de chegadas dos clientes à fila, o tempo de serviço, o número de servidores, a disciplina de serviço, a capacidade do sistema, etc.. Dado o seu enorme potencial em diversas áreas, a teoria de filas de espera tem sido, e continua a ser, muito estudada por matemáticos, sobretudo os que têm interesse em teoria da probabilidade e suas aplicações.

Esta dissertação é dedicada ao estudo de vários modelos de filas de espera e à apresentação de algumas das suas aplicações na área dos serviços de saúde que, tipicamente, são confrontados com atrasos no atendimento de pacientes.

O trabalho está divido em três grandes partes: uma primeira parte com conceitos básicos da teoria das probabilidades e de processos estocásticos, com especial destaque para as cadeias de Markov; na segunda parte descrevem-se alguns modelos de filas de espera, com especial destaque para os markovianos, e são explorados alguns aspetos relevantes de ponto de vista das aplicações como, por exemplo, o número médio de clientes no sistema (ou na fila), o tempo médio que um cliente aguarda no sistema (ou na fila) etc.; por fim são apresentadas aplicações concretas de filas de espera na área dos serviços de saúde através da exploração de alguns artigos em revistas internacionais como o *Journal of Medical Systems, European Journal of Operational Research* e *Health Care Management Sciences*.

Palavras-chave: Teoria das Probabilidades, Processos Estocásticos, Cadeias de Markov, Filas de Espera, Aplicações em Serviços de Saúde.

Abstract

Queueing theory is a branch of applied probability and operational research. This theory deals with the study of different probabilistic models used to describe the evolution of a queue. Aspects of interest in a queueing model are, for example: the arrival process of clients, the distribution of the service times, the number of servers, the discipline, the capacity of the system, etc. Given its huge potential in many areas, queuing theory has been, and continues to be, extensively studied by mathematicians, especially by those interested in probability theory and its applications.

In this dissertation, several models for queues are studied. Also, some applications of queuing models to health care services are presented. These type of services are usually affected by delays and long waiting times for patients.

The work is divided into three main parts: a first part contains basic concepts of probability theory and stochastic processes, with special emphasis on Markov chains; in the second part some queueing models are presented, with special emphasis for the markovian ones, and some relevant aspects for the applications point of view are presented, such as the expected number of clients in the system (or in the queue), the mean time a customer waits in the system (or queue) etc.; finally some applications of queueing theory in health services are explored in some articles published in international journals, such as Journal of Medical Systems, European Journal of Operational Research and Health Care Management Sciences.

Keywords: Probability Theory, Stochastic Processes, Markov Chains, Queues, Applications in Health Services.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

As filas de espera propriamente ditas fazem parte do dia-a-dia dos indivíduos na sociedade moderna. Nas nossas deslocações diárias, nos bancos, nos hospitais, nas repartições, nos cafés, no cinema etc., enfrentamos filas de espera. Segundo Hillier e Lieberman [18], nos Estados Unidos, estimou-se que os americanos gastam 37.000.000.000 horas por ano em esperas nas filas. Se este tempo fosse gasto produtivamente em vez disso, seria uma quantidade de quase 20 milhões pessoas por ano de trabalho útil.

Segundo [11], o primeiro estudo das filas de espera foi realizado pelo matemático dinamarquês A.K Erlang no ano 1909, onde Erlang utilizou esta teoria para resolver um problema de congestionamento de linhas telefónicas na Dinamarca. Erlang é considerado por muitos autores como o pai da teoria de filas de espera, devido ao facto de o seu trabalho se ter antecipado por várias décadas aos conceitos modernos desta teoria. Em 1917, Erlang publicou um trabalho com o título: "*Solutions of Some Problems in the Theory of Probabilities of Singnificance in Autmatic Telephone Exchange*", onde a sua experiência ficou reconhecida.

Desde então, as áreas de telecomunicações, da ciência da computação, de economia, da saúde, da administração e de processamentos de fluxos usufruíram dessa teoria. Entre os vários problemas estudados nestas áreas, destacam-se problemas de congestionamento de tráfego de pessoas, de escoamento de fluxos de carga em terminais, de carregamento ou descarregamento de produtos etc..

A teoria de filas de espera é um ramo da probabilidade aplicada e da investigação operacional. Segundo Preater [41], esta teoria preocupa-se em explicar por meio da modelação matemática, o fenómeno de congestionamento em sistemas de serviço. Esta teoria tem como objetivo principal optimizar o desempenho de um sistema, de modo a reduzir os seus custos operacionais e aumentar a satisfação do cliente [17].

Um dos objetivos deste trabalho foi explorar a aplicação da teoria das filas de espera à área da saúde. A escolha desta área deveu-se ao facto de existir uma ampla gama de serviços de saúde onde a formação de filas de espera é quase inevitável e pode ter efeitos muito prejudiciais à população que a eles recorre. Segundo [43], as aplicações no campo da saúde podem incluir a análise de filas de espera em quaisquer instalações e podem incidir sobre espaços, equipamentos e/ou pessoal.

Em geral, aplicação da teoria de filas de espera na área da saúde tem por objetivo

fornecer informações relevantes aos que analisam ou gerem o serviço de saúde, de modo a conseguir um equilíbrio apropriado entre os custos do serviço e o tempo de espera dos pacientes.

De salientar que o tema teoria de filas de espera é muito vasto e que seria impossível abordar nesta dissertação todos os modelos existentes na literatura. Por este motivo, aqui serão estudados essencialmente os modelos markovianos, mas também alguns não markovianos mais utilizados nas aplicações. Estaremos interessados em analisar, entre outras coisas, características como o número médio de clientes que aguardam na fila (ou no sistema), o tempo médio que um cliente aguarda na fila (ou no sistema), a probabilidade de um cliente ter que esperar, etc..

Para além desta Introdução, esta dissertação contém mais cinco capítulos.

O Capítulo 2 é dedicado apresentação de alguns conceitos básicos de teoria das probabilidades que são importantes para uma melhor compreensão do Capítulo 3, onde são abordados os processos estocásticos. Neste último capítulo é dado especial destaque às cadeias de Markov e ao processo de Poisson, que é utilizado em vários modelos de filas de espera.

No Capítulo 4 estudam-se vários modelos de filas de espera. É dada particular atenção a modelos markovianos, que fazem uso do processo de Poisson, mas também são ligeiramente abordados modelos não markovianos. Explora-se o comportamento dos modelos ao longo do tempo e estudam-se algumas das suas características como: o número médio de clientes na fila e no sistema, o tempo médio que um cliente aguarda na fila e no sistema, a probabilidade de um cliente ter que esperar, a probabilidade de um cliente abandonar o sistema (só para alguns modelos), etc..

No Capítulo 5 exploram-se algumas aplicações na área dos serviços de saúde de modelos estudados no Capítulo 4. Após uma pesquisa intensa em várias revistas científicas, selecionaram-se alguns artigos para serem estudados com mais pormenor, tentando abranger o maior número de modelos de filas de espera abordados nesta dissertação. Finalmente, no Capítulo 6 apresentam-se algumas conclusões e propostas de trabalho futuro.

Capítulo 2

Tópicos de Teoria de Probabilidades

Neste capítulo, serão enunciados alguns conceitos importantes da teoria das probabilidades, que serão essenciais para realizar uma análise teórica de um processo estocástico, em geral, e de uma fila de espera, em particular.

2.1 Espaços mensuráveis e definição axiomática de probabilidade

Um fenómeno aleatório é um fenómeno em que não é possível, a partir do passado, prever com exatidão o seu futuro. As experiências, cujos resultados são imprevisíveis, associadas a tais fenómenos aleatórios dizem-se experiências aleatórias. Supõe-se que tais experiências podem ser realizadas uma infinidade de vezes, sempre nas mesmas condições, chamando-se prova a cada uma das suas repetições ou realizações. O resultado da prova é um elemento imprevisível pertencente ao conjunto dos resultados possíveis da experiência em causa [13]. A teoria de probabilidades tem por objetivo construir um modelo matemático que permita descrever tais experiências aleatórias.

O conjunto de todos os resultados possíveis de uma experiência aleatória é chamado de *espaço amostral* e é usualmente representado por Ω. Um *acontecimento* é um subconjunto A do espaço amostral Ω. Os *acontecimentos elementares* são os que correspondem a subconjuntos singulares de Ω. Diz-se que se realiza o acontecimento A quando o resultado da experiência é algum $\omega \in A$.

Exemplo 2.1.1. *Um exemplo tradicional de experiência aleatória é o lançamento de um dado, com as faces numeradas de 1 a 6, sendo o resultado o número da face voltada para cima. Para esta experiência, o espaço amostral é* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Podemos enunciar alguns acontecimentos como:*

- *Ao acontecimento "sair face par"corresponde o subconjunto* $A = \{2, 4, 6\}$.
- *Ao acontecimento "não sair múltiplo de 5"corresponde o subconjunto* $B = \{1, 2, 3, 4, 6\}$.

- *Ao acontecimento "sair face par e múltiplo de 5"corresponde* $A \cap \overline{B} = \emptyset$.
- *Ao acontecimento "sair face ímpar ou múltiplo de 5"corresponde* $\overline{A} \cup \overline{B} = \{1, 3, 5\}$.
- *Ao acontecimento "sair face ímpar e múltiplo de 5"corresponde* $\overline{A} \cap B = \{5\}$.

Observe-se que este último acontecimento é elementar.

Definição 2.1.1. *Seja* Ω *um conjunto não vazio e* $\mathcal{P}(\Omega)$ *o conjunto das partes de* Ω. $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ *diz-se uma* σ*-álgebra sobre* Ω *se satisfaz os axiomas seguintes:*

i. $\Omega \in \mathcal{F}$.

ii. Se $A \in \mathcal{F}$, *então* $\overline{A} = \{\omega \in \Omega : \omega \notin A\} \in \mathcal{F}$;

iii. Se $\{A_n\}_{n\in\mathbb{N}}$ *é uma sucessão qualquer de elementos de* $\mathcal{F}$, *então* $\bigcup\limits_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

O conjunto Ω *munido de uma* σ*-álgebra* $\mathcal{F}$, *ou o par* $(\Omega, \mathcal{F})$, *chama-se espaço mensurável. Um elemento de* $\mathcal{F}$ *designa-se por conjunto mensurável.*

Numa σ*-álgebra* são válidas as seguintes propriedades:

P1. $\emptyset \in \mathcal{F}$;

P2. Se $\{A_n\}_{n\in\mathbb{N}}$ é uma sucessão qualquer de elementos de $\mathcal{F}$, então $\bigcap\limits_{n\in\mathbb{N}} A_n \in \mathcal{F}$;

P3. Se $A_i \in \mathcal{F}, i = 1, 2, ..., k,$ então $\bigcup_{i=1}^{k} A_i \in \mathcal{F}$ e $\bigcap_{i=1}^{k} A_i \in \mathcal{F}$.

Uma das $\sigma-$álgebras mais importante na teoria das probabilidades é a chamada $\sigma-$álgebra de Borel sobre $\mathbb{R}$.

Definição 2.1.2. *A* σ*-álgebra de Borel sobre* $\mathbb{R}$, *denota-se por* $\mathcal{B}$, *é a* σ*-álgebra gerada pelos intervalos reais da forma* $(-\infty, a], a \in \mathbb{R}$, *isto é, a menor* σ*-álgebra sobre* $\mathbb{R}$ *que contém todos estes intervalos.*

Nota: A $\sigma-$álgebra $\mathcal{B}$ pode ser também obtida através de outros conjuntos geradores, como por exemplo, os intervalos da forma $(a, b], a, b \in \mathbb{R}(a < b)$.

Definição 2.1.3. *Dados* $(\Omega_1, \mathcal{F}_1)$ *e* $(\Omega_2, \mathcal{F}_2)$ *espaços mensuráveis, uma função* $f : \Omega_1 \to \Omega_2$ *é dita mensurável se a imagem inversa de qualquer conjunto mensurável é mensurável, ou seja,*

$$A \in \mathcal{F}_2 \Rightarrow f^{-1}(A) \in \mathcal{F}_1.$$

Definição 2.1.4. *Considere um espaço mensurável* $(\Omega, \mathcal{F})$. *Chama-se medida de probabilidade sobre* $(\Omega, \mathcal{F})$ *a uma aplicação* $P : \mathcal{F} \to [0, \infty)$ *que verifica os seguintes axiomas:*

i. $P(\emptyset) = 0$;

ii. $P(\Omega) = 1$;

iii. Se $\{A_n\}_{n\in\mathbb{N}}$ *é uma sucessão de elementos de* $\mathcal{F}$, *com* $A_i \cap A_j = \emptyset (i \neq j)$, *então,*

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P\left(A_n\right).$$

Se P *é uma medida de probabilidade sobre* $(\Omega, \mathcal{F})$, *ao triplo* $(\Omega, \mathcal{F}, P)$ *é chamado de espaço de probabilidade.*

Teorema 2.1.1. *Seja* $(\Omega, \mathcal{F}, P)$ *um espaço de probabilidade. As seguintes propriedades são válidas:*

a) Se $A, B \in \mathcal{F}$ *e são tais que* $A \subseteq B$, *então* $P(A) \leq P(B)$ *e*

$$P(B \cap \overline{A}) = P(B) - P(A);$$

b) Para todo o $A \in \mathcal{F}$, *tem-se* $0 \leq P(A) \leq 1$ *e* $P(\overline{A}) = 1 - P(A)$;

c) Para todo $A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

d) Se $\{A_n\}_{n\in\mathbb{N}}$ *é uma sucessão crescente de elementos de* $\mathcal{F}$ *(isto é,* $A_n \subseteq A_{n+1}, n \in \mathbb{N}$*), então*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{n\in\mathbb{N}} A_n\right);$$

e) Se $\{A_n\}_{n\in\mathbb{N}}$ *é uma sucessão decrescente de elementos de* $\mathcal{F}$*(isto é,* $A_n \supseteq A_{n+1}, n \in \mathbb{N}$*), então*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcap_{n\in\mathbb{N}} A_n\right).$$

Um modelo matemático para uma experiência aleatória é um triplo $(\Omega, \mathcal{F}, P)$ em que Ω é o espaço amostral, $\mathcal{F}$ é a $\sigma-$álgebra que contém todos os acontecimentos de interesse e P a medida de probabilidade sobre $(\Omega, \mathcal{F})$.

2.2 Probabilidade condicional e independência

Sejam A e B dois acontecimentos de uma experiência aleatória modelada por um espaço de probabilidade $(\Omega, \mathcal{F}, P)$. Quando queremos calcular a probabilidade de ocorrer o acontecimento A sabendo que B ocorreu, estamos a querer calcular uma probabilidade condicional.

Definição 2.2.1. *Se* $P(B) > 0$, *então a probabilidade condicional que o acontecimento* A *ocorre dado que o acontecimento* B *ocorreu é definida por*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Definição 2.2.2. *Os acontecimentos A e B de um mesmo espaço de probabilidade $(\Omega, \mathcal{F}, P)$ dizem-se independentes se*

$$P(A \cap B) = P(A)P(B).$$

De um modo geral, se $\{A_n\}_{n\in\mathbb{N}}$ é uma sucessão qualquer de acontecimentos de um mesmo espaço de probabilidade $(\Omega, \mathcal{F}, P)$, diz-se que os acontecimentos são independentes se

$$\forall k \in \mathbb{N}, \forall \{B_1, ..., B_k\} \subset \{A_n\}_{n\in\mathbb{N}},\ P\left(\bigcap_{i=1}^{k} B_i\right) = \prod_{i=1}^{k} P(B_i).$$

Os acontecimentos que não são independentes são ditos ser dependentes.

Observação: Se A e B são independentes então $P(B|A) = P(B)$ e $P(A|B) = P(A)$, sempre que as probabilidades condicionais estejam definidas.

Teorema 2.2.1. *Seja $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade e $\{A_n\}_{n\in\mathbb{N}}$ uma partição do espaço amostral Ω, isto é, uma família $\{A_n\}_{n\in\mathbb{N}}$ de acontecimentos tais que $\bigcup_{n\in\mathbb{N}} A_n = \Omega$ e $A_i \cap A_j = \emptyset,\ i \neq j$. Se $P(A_n) > 0, n \in \mathbb{N}$, B e C dois acontecimentos então temos:*

- *[Teorema da probabilidade total]*

$$P(B) = \sum_{n\in\mathbb{N}} P(B|A_n)P(A_n);$$

- *[Teorema de Bayes] se $P(B) > 0$,*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum\limits_{n\in\mathbb{N}} P(A_n)P(B|A_n)},\ i \in \mathbb{N};$$

- *se $P(A_n \cap C) > 0,\ n \in \mathbb{N}$,*

$$P(B|C) = \sum_{n\in\mathbb{N}} P(B|A_n \cap C)P(A_n|C).$$

2.3 Variável aleatória e função de distribuição

As variáveis aleatórias têm uma importância fundamental na análise estatística de um fenómeno aleatório. Em geral, o que contamos ou medimos resulta da atribuição de valores numéricos aos resultados possíveis de uma experiência aleatória. Assim, quando estamos interessados em estudar uma ou mais caraterísticas numéricas relativas a uma experiência aleatória estamos a trabalhar com uma ou mais variáveis aleatórias.

Definição 2.3.1. *Considere-se um espaço de probabilidade* $(\Omega, \mathcal{F}, P)$ *e o espaço mensurável* $(\mathbb{R}, \mathcal{B})$*. Uma variável aleatória* X *é uma aplicação mensurável entre os conjuntos* Ω *e* $\mathbb{R}$*, ou seja,*

$$X : \Omega \longrightarrow \mathbb{R} \ \textit{tal que} \ X^{-1}(B) \in \mathcal{F}, \forall B \in \mathcal{B}.$$

Definição 2.3.2. *A função de distribuição de uma variável aleatória* X *é a função* $F : \mathbb{R} \longrightarrow [0,1]$ *dada por* $F(x) = P(X \leq x) = P(X \in]-\infty, x])$.

A função de distribuição F verifica as seguintes propriedades:

a) A função F é não decrescente e continua à direita.

b) $\lim_{x \to +\infty} F(x) = 1$ e $\lim_{x \to -\infty} F(x) = 0$.

c) $F(b) - F(a) = P(a < X \leq b) = P(X \in]a, b]), \forall a, b \in \mathbb{R}, a < b$.

Observações:

i. Se X e Y são duas variáveis aleatórias, diz-se que são identicamente distribuídas, denota-se por $(X \stackrel{d}{=} Y)$, se $F_X = F_Y$.

ii. A função de distribuição carateriza a variável aleatória, no sentido em que, se duas variáveis aleatórias X e Y têm a mesma função de distribuição, então

$$P(X \in B) = P(Y \in B), \ \forall B \in \mathcal{B}.$$

Ao longo deste trabalho serão utilizados dois tipos de variáveis aleatórias: as discretas e as absolutamente contínuas.

Definição 2.3.3. *A variável aleatória* X *diz-se discreta se existir um conjunto* $S \in \mathcal{B}$*, finito ou infinito numerável, tal que* $P(X \in S) = 1$*. Ao menor elemento de* $\mathcal{B}$ *que satisfaz esta condição chamamos suporte de* X*. Uma variável aleatória discreta é caraterizada pela chamada pela função de probabilidade que é dada por:*

$$f(x) = \begin{cases} P(X = x), & x \in S. \\ 0, & c.c, \end{cases}$$

A função de distribuição de uma variável aleatória discreta X é dada por

$$F(x) = \sum_{x_k \in S : x_k \leq x} P(X = x_k) = \sum_{x_k \in S : x_k \leq x} f(x_k),$$

em que S é o suporte de X e f é função de probabilidade de X.

Nota: Observe-se que se duas variáveis aleatórias discretas têm a mesma função de probabilidade, então elas têm a mesma função de distribuição. Deste modo, a função de probabilidade de uma variável discreta carateriza essa mesma variável.

Se as variáveis aleatórias discretas são caraterizadas à custa de um função de probabilidade, as absolutamente contínuas serão caraterizadas através de um função densidade de probabilidade.

Definição 2.3.4. *Uma função $f : \mathbb{R} \to \mathbb{R}$ é uma função densidade de probabilidade sobre $\mathbb{R}$ se satisfaz as seguintes condições:*

$$\begin{cases} f(x) \geq 0; \\ f \text{ é integravel e } \int_{-\infty}^{+\infty} f(x)dx = 1. \end{cases}$$

Definição 2.3.5. *A variável aleatória X diz-de absolutamente contínua se existir f, uma função densidade de probabilidade sobre $\mathbb{R}$, tal que*

$$P(X \in B) = \int_B f(x)dx, \ \forall B \in \mathcal{B}.$$

Tal função f é chamada de função densidade de probabilidade de X.

Se X é absolutamente contínua, com função densidade de probabilidade f, a função de distribuição de X é dada por

$$F(x) = \int_{-\infty}^{x} f(t)dt, \ x \in \mathbb{R}.$$

Nota: Observe-se que, se duas variáveis aleatórias absolutamente contínuas têm a mesma função de probabilidade, então elas têm a mesma função de distribuição. Deste modo, a função densidade de probabilidade de uma variável absolutamente contínua carateriza essa mesma variável.

2.3.1 Valor esperado e variância de uma variável aleatória

Nesta secção, vamos apresentar medidas de localização e de dispersão de uma variável aleatória. Em particular, iremos apresentar a definição de valor esperado (medida de localização), de variância, desvio padrão e coeficiente de variação (medidas de dispersão) para o caso discreto e para o caso absolutamente contínuo. Serão ainda apresentadas as definições de função geradora de momentos e de função geradora de probabilidades, que se obtêm à custa do valor esperado de certas funções da variável aleatória em causa.

Definição 2.3.6. *Dada uma variável aleatória X, o valor esperado (ou valor médio) de X é denotado por $E[X]$ e é dado por:*

- *No caso de X ser discreta, com suporte S e função de probabilidade f,*

$$E[X] = \sum_{x_k \in S} x_k f(x_k) \text{ se } \sum_{x_k \in S} |x_k| f(x_k) < \infty.$$

  Se $\sum_{x_k \in S} |x_k| f(x_k) = \infty$, não existe valor esperado.

- *No caso de X ser absolutamente contínua, com função densidade de probabilidade f,*

$$E(X) = \int_{-\infty}^{+\infty} x f(x)dx \text{ se } \int_{-\infty}^{+\infty} |x| f(x)dx < \infty.$$

  Se $\int_{-\infty}^{+\infty} |x| f(x)dx = \infty$, não existe valor esperado.

Em muitas situações estaremos interessados em calcular o valor esperado de uma função de uma variável aleatória X. Tal será feito de acordo com a definição seguinte.

Definição 2.3.7. *Seja X uma variável aleatória e $H : D \subseteq \mathbb{R} \to \mathbb{R}$ tal que $H(X)$ ainda é uma variável aleatória. O valor esperado de $H(X)$ é dado por:*

- *No caso de X ser discreta, com suporte S e função de probabilidade f,*

$$E[H(X)] = \sum_{x_k \in S} H(x_k) f(x_k) \text{ se } \sum_{x_k \in S} |H(x_k)| f(x_k) < \infty.$$

  Se $\sum_{x_k \in S} |H(x_k)| f(x_k) = \infty$, não existe valor esperado.

- *No caso de X ser absolutamente contínua, com função densidade de probabilidade f,*

$$E[H(X)] = \int_{-\infty}^{+\infty} H(x) f(x) dx \text{ se } \int_{-\infty}^{+\infty} |H(x)| f(x) dx < \infty.$$

  Se $\int_{-\infty}^{+\infty} |H(x)| f(x) dx = \infty$, não existe valor esperado.

Estamos agora em condições de definir a variância, o desvio padrão e o coeficiente de variação de uma variável aleatória.

Definição 2.3.8. *Dada uma variável aleatória X,*

- *a variância de X é denotada por $Var[X]$ e é dada por*

$$Var[X] = E[(X - E[X])^2],$$

  desde que tal valor esperado exista. Quando este valor esperado não existe diz-se que não existe variância.

- *quando a variância de X existe, define-se desvio padrão de X como sendo a raiz quadrada da variância, isto é, $\sigma = \sqrt{Var[X]}$.*

- *quando o valor esperado de X for diferente de zero e existir variância de X, define-se o coeficiente de variação de X como sendo o quociente entre o desvio padrão e o valor esperado, isto é,*

$$CV = \frac{\sigma}{E[X]}.$$

De seguida vamos enunciar algumas propriedades do valor esperado e da variância.

Propriedades do valor esperado e da variância: Sejam c e d constantes reais e sejam X e Y variáveis aleatórias. Tem-se:

1) $E[c] = c$ e $Var[c] = 0$;

2) se existir $E[X]$, então $E[cX] = cE[X]$, e, se existir $Var[X]$, então

$$Var[cX] = c^2 Var[X];$$

3) Se existirem $E[X]$ e $E[Y]$, então $E[X + Y] = E[X] + E[Y]$;

4) Se existir $Var[X]$, então $Var[X + d] = Var[X]$.

Apresentamos de seguida as definições de função geradora de momentos e de função geradora de probabilidades. Estas funções obtêm-se a custa do valor esperado de certas funções da variável em causa.

Definição 2.3.9. *Seja X uma variável aleatória.*

- *A função geradora de momentos de X é a função M definida por*

$$M(t) = E[e^{tX}],$$

para os valores de $t \in \mathbb{R}$ em que $E[e^{tX}]$ existe.

- *Se X for discreta com suporte $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$, a função geradora de probabilidades de X é a função definida por*

$$G(z) = E[z^X],\ |z| \leq 1.$$

Observações:

1) Estas funções geradoras caraterizam a variável em causa. De facto, se duas variáveis aleatórias têm a mesma função geradora de momentos, elas terão a mesma função de distribuição. Analogamente, se duas variáveis aleatórias discretas de suporte $\mathbb{N}_0$ tiverem a mesma função geradora de probabilidades, então elas terão a mesma função de distribuição.

2) O momento de ordem k, $k \in \mathbb{N}$, da variável aleatória X, definido por $E[X^k]$, pode obter-se a custa da derivada de ordem k da função geradora de momentos calculada em zero, isto é,

$$M^{(k)}(0) = E[X^k],$$

quando tal derivada existir.

3) Através da derivada de ordem k, $k \in \mathbb{N}_0$, da função geradora de probabilidades calculado em zero, podemos obter a função de probabilidade da variável em causa. De facto,

$$P(X = k) = G^{(k)}(0),\ k \in \mathbb{N}_0.$$

2.3.2 Variáveis aleatórias independente

A independência entre variáveis aleatórias significa que o conhecimento do resultado obtido para uma delas não altera a probabilidade de ocorrência dos diferentes resultados das outras. Para sermos mais precisos apresentamos a definição a seguir.

Definição 2.3.10. *Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias todas definidas sobre o mesmo espaço de probabilidade $(\Omega, \mathcal{F}, P)$. Estas variáveis são independentes se para todos os conjuntos $B_1 \in \mathcal{B}, ..., B_n \in \mathcal{B}$,*

$$P\left(\bigcap_{i=1}^{n}(X_i \in B_i)\right) = \prod_{i=1}^{n} P(X_i \in B_i).$$

Observacao: Se as variáveis aleatórias $X_1, X_2, ..., X_n$, além de independentes, tiverem a mesma função de distribuição, diz-se que são independentes e identicamente distribuídas e abrevia-se por *i.i.d.*

Para terminar esta secção vamos enunciar algumas propriedades válidas para variáveis aleatórias independentes.

Propriedades: Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes.

i. Se $g_i : D_i \subseteq \mathbb{R} \to \mathbb{R},\ i = 1, ..., n$, são tais que $g_i(X_i),\ i = 1, ..., n$, ainda são variáveis aleatórias, então

$$g_1(X_1), ..., g_n(X_n)$$

também são independentes.

ii. Se existirem os valores esperados e as variâncias, então

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i], \text{ e } Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var[X_i].$$

iii. Seja $Y = min\{X_1, X_2, ..., X_n\}$. A função de distribuição de Y é:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= 1 - P(Y > y) \\ &= 1 - P\left((X_1 > y) \cap ... \cap (X_n > y)\right) \\ &= 1 - \prod_{i=1}^{n} P(X_i > y) \\ &= 1 - \prod_{i=1}^{n} (1 - F_{X_i}(y)), \end{aligned}$$

onde F_{X_i} representa a função de distribuição de X_i. Note-se que na penúltima igualdade usou-se a independência das variáveis.

Observação: Se adicionalmente as variáveis forem identicamente distribuídas, isto é, tiverem a mesma função de distribuição F, então a função de distribuição de Y reduz-se a

$$F_Y(y) = 1 - (1 - F(y))^n. \tag{2.1}$$

iv. Seja $Y = max\{X_1, X_2, ..., X_n\}$. A função de distribuição de Y é

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\left((X_1 \leq y) \cap ... \cap (X_n \leq y)\right) \\ &= \prod_{i=1}^{n} P(X_i \leq y) \\ &= \prod_{i=1}^{n} F_{X_i}(y). \end{aligned}$$

Na penúltima igualdade usou-se o mesmo procedimento da propriedade anterior.

Observação: Se adicionalmente as variáveis forem identicamente distribuídas, isto é, tiverem a mesma função de distribuição F, então a função de distribuição de Y reduz-se a

$$F_Y(y) = (F(y))^n.$$

2.3.3 Algumas distribuições mais utilizadas

De seguida definimos três distribuições que terão grande importância em processos estocásticos em geral, e em modelos de filas de espera, em particular, abordados nos capítulos 3 e 4 deste trabalho respetivamente.

Definição 2.3.11. *Diz-se que uma variável aleatória discreta X tem distribuição de Poisson com parâmetro λ, $\lambda \in \mathbb{R}^+$, se a sua função de probabilidade for dada por*

$$f(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!}, & x = 0, 1, 2, ... \\ 0, & c.c \end{cases}.$$

Uma demonstração simples (ver [44], pag.38) permite concluir que,

$$E[X] = \ Var[X] = \lambda.$$

Definição 2.3.12. *Diz-se que uma variável aleatória absolutamente contínua X tem distribuição exponencial com parâmetro λ, $\lambda \in \mathbb{R}^+$, se a sua função densidade de probabilidade é dada por*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & para \quad x \geq 0 \\ 0, & para \quad x < 0 \end{cases}.$$

Notas: Se X tem distribuição exponencial de parâmetro λ então:

1) A sua função de distribuição é

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\lambda x}, & para \quad x \geq 0 \\ 0, & para \quad x < 0 \end{cases}.$$

2) Uma demonstração simples (ver [44], pag.39) permite concluir que o valor esperado e a variância de X são dados por

$$E[X] = \frac{1}{\lambda} \text{ e } Var[X] = \frac{1}{\lambda^2}.$$

3) X tem a chamada propriedade de falta de memória, isto é,

$$P(X > t + s | X > s) = P(X > t), \ s, t \in \mathbb{R}_0^+. \tag{2.2}$$

A demonstração desta propriedade faz-se sem nenhuma dificuldade, pois basta recorrer à função distribuição da variável X.
De referir que, entre as variáveis absolutamente contínuas, apenas as que têm distribuição exponencial possuem esta propriedade. De facto, se X é uma variável aleatória absolutamente contínua que tem a propriedade de falta de memória e sendo $G(x) = P(X > x)$, então, $\forall s, t \geq 0$, tem-se:

$$\begin{array}{ccc} P(X > t + s | X > s) = P(X > t) & \Leftrightarrow & \frac{P(X>t+s, X>s)}{P(X>s)} = P(X > t) \Leftrightarrow \\ \frac{P(X>t+s)}{P(x>s)} = P(X > t) & \Leftrightarrow & G(t + s) = G(s)G(t). \end{array}$$

Concluímos assim que G é uma função que satisfaz a conhecida equação funcional de Cauchy, $H(t+s) = H(s)H(t)$. É sabido que as únicas funções contínuas à direita que satisfazem esta condição são funções do tipo $H(t) = e^{-\lambda t}$. Uma vez que a nossa função G é contínua à direita, então a função de distribuição de X tem expressão dada por $F(x) = 1 - e^{-\lambda x}$.

Definição 2.3.13. *Diz-se que uma variável aleatória absolutamente contínua X tem distribuição Gama com parâmetros α e β se a sua função densidade de probabilidade é dada por*

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & x \leq 0 \end{cases},$$

onde $\alpha, \beta \in \mathbb{R}^+$ e Γ é chamada função gama definida por

$$\Gamma(\alpha) = \int_0^{+\infty} e^{-x} x^{\alpha-1} dx.$$

Notas: Se X tem distribuição Gama com parâmetros α e β então:

1) Uma demonstração simples (ver [44], pag.39) permite concluir que o valor esperado e a variância de X são dados por

$$E[X] = \frac{\alpha}{\beta} \text{ e } Var[X] = \frac{\alpha}{\beta^2}.$$

2) Quando $\alpha = 1$, X tem distribuição exponencial com parâmetro β.

3) Se $X_1, ..., X_n$ são variáveis aleatórias independentes e tais que X_i tem distribuição Gama com parâmetros α_i e β, $i \in \{1, ..., n\}$, então a soma de tais variáveis tem distribuição Gama com parâmetros $\sum_{i=1}^{n} \alpha_i$ e β. A demonstração deste resultado faz uso da função geradora de momentos e da independência das variáveis, como se poderá ver de seguida.

Demonstração: Comecemos por determinar a função geradora de momentos de uma variável X com distribuição Gama com parâmetros α e β. A função geradora de momentos de X é dado por:

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \int_0^{+\infty} e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} x^{\alpha-1} e^{-(\beta-t)x} dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} \left(\frac{y}{\beta-t}\right)^{\alpha-1} e^{-y} \frac{1}{\beta-t} dy \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(\beta-t)^\alpha} \Gamma(\alpha) \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha, \text{ se } \beta - t > 0 \Leftrightarrow t < \beta. \end{aligned}$$

Na primeira igualdade utilizou-se a definição do valor esperado de uma função da variável X e na antepenúltima usou-se a substituição $y = (\beta - t)x$.

Considera-se agora que $X_1, ..., X_n$ são variáveis aleatórias independentes, tais que X_i tem distribuição Gama com parâmetros α_i e β, $i \in \{1, ..., n\}$, e vamos determinar a função geradora de momentos da variável $Y = \sum_{i=1}^{n} X_i$. Tem-se que:

$$\begin{aligned} M_Y(t) &= E\left[e^{t\sum_{i=1}^n X_i}\right] \\ &= E\left[\prod_{i=1}^{n} e^{tX_i}\right] \\ &= \prod_{i=1}^{n} E[e^{tX_i}] \\ &= \prod_{i=1}^{n} \left(\frac{\beta}{\beta-t}\right)^{\alpha_i} \\ &= \left(\frac{\beta}{\beta-t}\right)^{\sum_{i=1}^n \alpha_i}, \end{aligned}$$

que coincide com a função geradora de momentos de uma distribuição Gama com parâmetros $\sum_{i=1}^{n} \alpha_i$ e β, como se pretendia mostrar. Note-se que na terceira igualdade se usou o facto de as variáveis aleatórias $e^{tX_1}, ..., e^{tX_n}$ serem independentes, uma vez que $X_1, ..., X_n$ também são independentes por hipótese.

4) A conjugação das duas notas anteriores permite-nos concluir que se $T_1, ..., T_n$ são variáveis aleatórias independentes e tais que $T_i,\ i \in \{1, ..., n\}$, tem distribuição exponencial com parâmetro β, então a soma de tais variáveis tem distribuição Gama com parâmetros n e β.

Observação: Nos modelos de filas de espera, a distribuição de Poisson será usada na descrição do processo de chegada de clientes ao sistema e a distribuição exponencial será usada na descrição dos tempos de chegada e de atendimento dos clientes. A distribuição Gama será usada para descrever a distribuição da soma dos tempos de atendimento de vários clientes (considerando que cada um dos tempos segue uma distribuição exponencial).

Capítulo 3

Processos Estocásticos

3.1 Introdução

Os processos estocásticos são utilizados para modelar a evolução de um fenómeno aleatório que varia ao longo do tempo. Assim, um processo estocástico será uma sucessão de variáveis aleatórias $\{X(t), t \in T\}$, todas definidas sobre um mesmo espaço de probabilidade $(\Omega, \mathcal{F}, P)$, e indexadas por um parâmetro t que, em geral, representa o tempo. Segue-se uma definição mais precisa.

Definição 3.1.1. *Considere-se um espaço de probabilidade $(\Omega, \mathcal{F}, P)$ e o espaço mensurável $(\mathbb{R}, \mathcal{B})$. Seja $T \neq \emptyset$ um conjunto e, para cada $t \in T$, tome-se a variável aleatória $X(t)$, função mensurável de $\Omega \longrightarrow \mathbb{R}$. A família de variáveis aleatórias $\{X(t), t \in T\}$ indexadas pelo parâmetro t chama-se processo estocástico definido sobre o espaço de probabilidade $(\Omega, \mathcal{F}, P)$. O conjunto de todos os valores que as variáveis $X(t)$ podem assumir é chamado espaço de estados do processo estocástico.*

Os processos estocásticos são classificados em função da natureza do conjunto de índices T e do espaço de estados S. Assim:

- Quando T é um conjunto finito ou infinito numerável, o processo estocástico é considerado um processo de tempo discreto. Neste caso, em geral, considera-se que $T = \{0, 1, 2 \dots\}$ e usa-se a notação $\{X_n,\ n \geq 0\}$ em vez de $\{X(t),\ t \in T\}$.
- Quando T é um conjunto infinito não numerável, o processo estocástico é considerado um processo de tempo contínuo. Neste caso, em geral, considera-se que $T = [0, +\infty[$ e usa-se a notação usual $\{X(t),\ t \geq 0\}$.
- Quando S é um conjunto finito ou infinito numerável, o processo é dito discreto.
- Quando S é um conjunto infinito não numerável, o processo é dito contínuo.

Entre os processos estocásticos existe um conjunto particularmente importante, que é muito estudado na literatura e que tem grandes aplicações práticas. Estamos a referir-nos aos processos que possuem a chamada propriedade de Markov, cuja definição é dada de seguida.

Definição 3.1.2. *Um processo de Markov* $\{X(t),\ t \in T\}$ *é um processo estocástico que tem a seguinte propriedade: dado que o valor da variável* $X(t)$ *é conhecido, o comportamento probabilístico das variáveis* $X(s)$*, com* $s > t$*, não é influenciado pelo conhecimento adicional dos valores das variáveis* $X(u)$*, com* $u < t$*.*

Por outras palavras, num processo de Markov, a probabilidade de qualquer comportamento futuro particular do processo, quando seu estado atual é conhecido, não é alterada pelo conhecimento adicional sobre o seu comportamento em instantes passados.

As definições mais concretas para o caso em que o tempo é discreto e para o caso em que o tempo é contínuo serão dadas nas secções seguintes.

3.2 Cadeia de Markov de tempo discreto

3.2.1 Definição e probabilidades de transição

Um processo de Markov de tempo discreto, $\{X_n,\ n \geq 0\}$, é designado de *cadeia de Markov de tempo discreto* se o correspondente espaço de estados S é um conjunto finito ou infinito numerável. Neste caso, a propriedade de Markov descreve-se do seguinte modo:

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_{n+1} = j | X_n = i), \qquad (3.1)$$

para todos os instantes de tempo, $n \in \mathbb{N}$, e todos os estados $i_0, ..., i_{n-1}, i, j$. Observe-se que esta propriedade diz-nos que a probabilidade de o processo estar no estado j no instante $n+1$, dado que conhecemos o trajeto do processo desde o instante inicial até ao instante n, depende somente do estado do processo no instante n, não dependendo de toda informação do passado do processo $(X_0, X_1, \ldots, X_{n-1})$.

Segundo (Grimmett e Stirzaker [16], pag.214), a evolução de uma cadeia de Markov de tempo discreto pode ser descrita através das chamadas probabilidades de transição do estado j para o estado i, isto é, pelas seguintes probabilidades:

$$p_{ij}^{n,n+1} \equiv P(X_{n+1} = j | X_n = i),\ i, j \in S,\ n \in \mathbb{N}_0. \qquad (3.2)$$

Tais probabilidades podem ser bastante complicadas, em geral, uma vez que dependem dos três elementos n, i e j. Neste trabalho, vamos considerar essencialmente o caso em que estas probabilidades não dependem de n, isto é, o caso das cadeias de Markov homogéneas.

Definição 3.2.1. *Uma cadeia de Markov* $\{X_n, n \geq 0\}$ *é dita homogénea se as suas probabilidades de transição não dependem de* n*, isto é,*

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i),\ \forall n \in \mathbb{N}_0,\ i, j \in S. \qquad (3.3)$$

Neste caso, usa-se apenas a notação p_{ij} *para denotar as probabilidades de transição, isto é, se*

$$p_{ij} \equiv P(X_1 = j | X_0 = i).$$

Numa cadeia de Markov homogénea, as probabilidades de transição são habitualmente apresentadas numa matriz $P = [p_{ij}]_{i,j\in S}$, quadrada de dimensão $\#S \times \#S$, chamada de matriz de probabilidade de transição. Assim, se a cadeia de Markov tiver espaço de estados $S = \{0, 1, 2, ...\}$, as respetivas probabilidades de transição p_{ij} são apresentadas numa matriz de dimensão infinita, ou seja,

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0j} & \cdots \\ p_{10} & p_{11} & \cdots & p_{1j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{i0} & p_{i1} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Naturalmente, quando S é finito a matriz $\mathbf{P}$ tem dimensão finita.

Observe que os elementos de uma matriz de transição $\mathbf{P}$ são probabilidades condicionais e, portanto, têm que satisfazer as seguintes condições:

1) $p_{ij} \geq 0$, para todo $i, j \in S$;

2) $\sum\limits_{j\in S} p_{ij} = 1$, para todo $i \in S$.

Observação: Esta última condição é facilmente demonstrada. De facto, para todo $i \in S$,

$$\sum_{j\in S} p_{ij} = \sum_{j\in S} P(X_1 = j|X_0 = i) = P\left(\bigcup_{j\in S}(X_1 = j)\middle| X_0 = i\right) = P(\Omega|X_0 = i) = 1,$$

uma vez que $\{(X_1 = j)\}_{j\in S}$ é uma sucessão de acontecimentos que forma uma partição de Ω e que uma probabilidade condicionada ainda é uma medida de probabilidade sobre $(\Omega, \mathcal{F})$.

Exemplo 3.2.1. *Considere que a chance de chover amanhã depende apenas das condições climáticas de hoje (se hoje está ou não a chover), e não depende de condições climáticas dos dias anteriores. Considere também que, se chover hoje, não choverá amanhã com probabilidade de* 0.7 *e que, se não chover hoje, então choverá amanhã com probabilidade de* 0.4*.*

Para construir a matriz das probabilidades de transição, $\boldsymbol{P}$*, primeiro devemos identificar o espaço de estados* S*. Se considerarmos que o processo está no estado 0 quando não chove e que está no estado* 1 *quando chove então, temos uma cadeia de Markov com dois estados e* $S = \{0, 1\}$*. Segundo, deve-se identificar todas as transições entre os estados. Neste caso temos* 4 *transições:*

$$0 \to 0, \;\; 0 \to 1, \;\; 1 \to 0, \;\; 1 \to 1$$

e as respetivas probabilidades de transição são:

$$p_{00} = 0.6, \;\; p_{01} = 0.4, \;\; p_{10} = 0.7, \;\; p_{11} = 0.3.$$

A matriz de probabilidades de transição é assim dada por

$$\boldsymbol{P} = \begin{pmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{pmatrix}.$$

As probabilidades referidas em (3.2), são designadas de probabilidades de transição de um passo. Interessa agora definir as probabilidades de transição de m passos, isto é,

$$P(X_{l+m} = j | X_l = i), l \in \mathbb{N}_0, m \in \mathbb{N}, i, j \in S. \tag{3.4}$$

É fácil perceber que, tal como acontece com as probabilidades de um passo, numa cadeia de Markov homogénea as probabilidades de transição de m-passos também não dependerão de l, pelo que basta considerarmos apenas as seguintes probabilidades

$$p_{ij}^{(m)} \equiv P(X_m = j | X_0 = i),\ m \in \mathbb{N}.$$

As **equações de Chapmam-Kolmogorov** fornecem um método adequado para calcular as probabilidades de transição de m-passos. Essas equações são estabelecidas pela seguinte fórmula:

$$p_{ij}^{(n+m)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)} \text{ para todos } n, m \in \mathbb{N} \text{ e todos } i, j \in S. \tag{3.5}$$

A demonstração destas equações faz-se sem grandes dificuldades. De facto,

$$\begin{aligned} p_{ij}^{(n+m)} &= P(X_{m+n} = j | X_0 = i) \\ &= \sum_{k \in S} P(X_{m+n} = j | X_n = k, X_0 = i) P(X_n = k | X_0 = i) \\ &= \sum_{k \in S} p_{kj}^{(m)} p_{ik}^{(n)}, \end{aligned}$$

tendo-se usado o ponto 3 do Teorema 2.2.1, na segunda igualdade, e a propriedade de Markov na terceira igualdade. Observe-se que a quantidade $p_{ik}^{(n)} p_{kj}^{(m)}$ representa a probabilidade de a cadeia passar do estado i para o estado j em $(n+m)$-passos através de um caminho que a leva ao estado k no n-ésimo passo.

Vamos agora considerar a matriz de probabilidades de transição de m-passos, isto é,

$$\mathbf{P}^{(m)} = \left[p_{ij}^{(m)} \right]_{i,j \in S},\ m \in \mathbb{N}. \tag{3.6}$$

As equações de Chapmam-Kolmogorov, estabelecidas em (3.5), podem agora ser escritas na forma matricial do seguinte modo

$$\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)} \mathbf{P}^{(m)},\ \ n, m \in \mathbb{N}, \tag{3.7}$$

sendo que $\mathbf{P}^{(1)}$ é a matriz de probabilidades de transição de um passo. No caso homogéneo, tem-se que

$$\mathbf{P}^{(n)} = \mathbf{P}^n,\ n \in \mathbb{N}, \tag{3.8}$$

sendo $\mathbf{P}$ a matriz de probabilidades de transição de um passo.

Observação: Esta igualdade mostra-se facilmente por indução. É obviamente válida para $n = 1$ e, se for válida para n, então:

$$\mathbf{P}^{(n+1)} = \mathbf{P}^{(n)}\mathbf{P}^{(1)} = \mathbf{P}^{n}\mathbf{P}^{1} = \mathbf{P}^{n+1},$$

é válida para $n+1$. Note-se que na primeira igualdade usamos a equação (3.7) e na segunda igualdade usamos a hipótese indução.

Para identificar a função de probabilidade da variável X_n, para além das probabilidades de transição, é necessário conhecer a função de probabilidade da variável X_0. Assim, se

$$\alpha_i \equiv P(X_0 = i),\ i \in S,$$

a função de probabilidade da variável X_n pode ser obtida do seguinte modo:

$$P(X_n = j) = \sum_{i \in S} P(X_n = j | X_0 = i) P(X_0 = i) = \sum_{i \in S} p_{ij}^{(n)} \alpha_i, \quad j \in S, \tag{3.9}$$

tendo sido utilizado na primeira igualdade o teorema da probabilidade total.

Exemplo 3.2.2. *(Muller [35], p. 119) " Um determinado indivíduo modifica o seu estado de espírito ao longo do seu dia de trabalho. Tendo sido observado pelos seus colegas durante um longo período, foram-lhe atribuídas as seguintes probabilidades de mudança do seu estado de espírito:*

- *se está de bom humor durante uma certa hora, a probabilidade de estar de mau humor durante a hora seguinte é de* 0.2*;*
- *se está de mau humor durante uma certa hora, a probabilidade de continuar de mau humor durante a hora seguinte é de* 0.4*.*

Pretende-se responder às seguintes questões:

a) se o indivíduo durante a primeira hora de trabalho estava de mau humor, qual é a probabilidade de estar de bom humor durante a terceira hora de trabalho?

b) Admitindo que os estados de espírito são igualmente prováveis quando o indivíduo chega ao trabalho, determine a probabilidade de ele estar de bom humor durante a terceira hora de trabalho?"

***Solução**: Considerando o estado do indivíduo como* 0 *quando está de bom humor e* 1 *quando está de mau humor, respetivamente, então temos o espaço de estados* $S = \{0, 1\}$ *e matriz de probabilidades de transição*

$$\boldsymbol{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix}.$$

a) Pretende-se calcular $P(X_3 = 0|X_1 = 1) = p_{10}^{(2)}$. *A partir da equação* (3.5) *vem,*

$$P_{10}^{(2)} = p_{10}p_{00} + p_{11}p_{10} = 0.72.$$

A probabilidade de o indivíduo estar de bom humor duas horas depois, sabendo que na primeira hora de trabalho esteve de mau humor, é 0.72.

A outra maneira de calcular a probabilidade $p_{10}^{(2)}$, *passa por obter a matriz* $\boldsymbol{P}^2$ *e localizar o elemento da linha 2 e coluna 1. Temos então*

$$\boldsymbol{P}^2 = \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \cdot \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.76 & 0.24 \\ 0.72 & 0.28 \end{pmatrix}$$

e $p_{10}^{(2)} = 0.72$.

b) Pretende-se calcular $P(X_3 = 0)$, *com* $\alpha_i = 0.5$, $i \in \{0, 1\}$. *A partir da equação* (3.9) *vem,*

$$P(X_3 = 0) = p_{00}^{(2)}0.5 + p_{10}^{(2)}0.5 = 0.744.$$

Assim a probabilidade de o indivíduo estar de bom humor durante a terceira hora de trabalho é 0.744.

3.2.2 Teoremas limite e distribuição estacionária

Nesta secção serão enunciados alguns resultados relativos ao comportamento limite de uma cadeia de Markov homogénea, sendo o limite estudado quando o tempo tende para infinito, isto é, quando $n \to \infty$. Vamos começar por apresentar a definição de distribuição limite.

Definição 3.2.2. *Seja* $\{X_n, n \geq 0\}$ *uma cadeia de Markov homogénea, com espaço de estado* S, *e* $(\pi_j,\ j \in S)$ *uma distribuição de probabilidade sobre* S, *isto é,* $\pi_j > 0$ *e* $\sum_{j\in S} \pi_j = 1$. *Diz-se que* $(\pi_j,\ j \in S)$ *é a distribuição limite da cadeia de Markov se para todo* $i, j \in S$ *se tem*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j,$$

sendo $p_{ij}^{(n)}$ *a probabilidade de transição de n-passos do estado i para o estado j.*

Observe-se que, quando a distribuição limite existe, ela coincide com o limite da distribuição da variável X_n, isto é, π_j coincide com $\lim_{n\to+\infty} P(X_n = j)$, e é independente da distribuição da variável X_0 . De facto, para todo $j \in S$, tem-se

$$\begin{aligned} \lim_{n\to\infty} P(X_n = j) &= \lim_{n\to\infty} \sum_{i\in S} P(X_n = j|X_0 = i)P(X_0 = i) = \sum_{i\in S} \left(\lim_{n\to\infty} p_{ij}^{(n)}\right) P(X_0 = i) \\ &= \pi_j \sum_{i\in S} P(X_0 = i) = \pi_j. \end{aligned}$$

Quando existe a distribuição limite de uma cadeia de Markov homogénea então ela pode ser obtida através da chamada distribuição estacionária da cadeia. Segue-se a definição de distribuição estacionária.

Definição 3.2.3. *Considere-se uma cadeia de Markov homogénea* $\{X_n, n \geq 0\}$*, com espaço de estado* S *e matriz de probabilidades de transição de um passo* $\mathbf{P} = [p_{ij}]_{i,j \in S}$*. Diz-se que uma distribuição de probabilidade* $\boldsymbol{\pi} = (p_j,\ j \in S)$ *sobre* S *é estacionária para esta cadeia de Markov se*

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P},\ \textit{isto é},\ p_j = \sum_{i \in S} p_i p_{ij},\ \forall j \in S.$$

A definição anterior estabelece uma relação entre a distribuição estacionária e matriz de probabilidades de transição de um passo. O teorema que se segue é uma consequência da definição e estabelece uma relação entre a distribuição estacionária e a matriz de probabilidades de transição de n-passos.

Teorema 3.2.1. *Seja* $\boldsymbol{\pi} = (p_j,\ j \in S)$ *uma distribuição estacionária para a cadeia de Markov homogénea* $\{X_n, n \geq 0\}$*, com espaço de estados* S*. Então, para todo* $n \in \mathbb{N}$*, tem-se*

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^n,\ \textit{isto é},\ p_j = \sum_{i \in S} p_i\, p_{ij}^{(n)},\ \forall j \in S.$$

A demonstração deste teorema faz-se facilmente por indução. Para $n = 1$, usa-se a definição de distribuição estacionária. Se o resultado for válido para n então também é válido para $n + 1$, uma vez que

$$\boldsymbol{\pi}\mathbf{P}^{n+1} = \boldsymbol{\pi}\mathbf{P}^n\mathbf{P} = \boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi},$$

tendo sido usada a hipótese indução, na segunda igualdade, e o facto de $\boldsymbol{\pi}$ ser uma distribuição estacionária, na última igualdade.

Deste teorema resulta uma consequência interessante para a situação em que a cadeia de Markov se inicia com uma distribuição que é estacionária. Observe-se que, nessa situação, tem-se: $\forall n \in \mathbb{N}$ e $\forall j \in S$,

$$P(X_n = j) = \sum_{i \in S} P(X_0 = i) P(X_n = j | X_0 = i) = \sum_{i \in S} p_i\, p_{ij}^{(n)} = p_j,$$

tendo sido usado o teorema anterior na última igualdade. Portanto, constata-se que, nessa situação, se a distribuição de X_0 for estacionária para a cadeia de Markov então as diferentes variáveis $X_1, X_2, ..., X_n, ...$ também têm a distribuição estacionária.

O teorema que se segue estabelece uma relação entre a distribuição limite de uma cadeia de Markov homogénea e uma distribuição estacionária para a mesma cadeia. A demonstração do teorema é longa e pode ser consultada em Müller ([35], pag. 101-104).

Teorema 3.2.2. *Considere-se uma cadeia de Markov homogénea* $\{X_n, n \geq 0\}$*, com espaço de estados* S*, e que tem distribuição limite* $(\boldsymbol{\pi}_j,\ j \in S)$*. Então* $(\boldsymbol{\pi}_j, j \in S)$ *constitui uma distribuição de probabilidade estacionária para a cadeia de Markov, ou seja,*

$$\boldsymbol{\pi}_j = \sum_{i \in S} \boldsymbol{\pi}_i\, p_{ij}.$$

Além disso, esta distribuição é a única distribuição estacionária para a cadeia de Markov.

Conjugando os dois últimos teoremas, demonstra-se o seguinte corolário.

Corolário 3.2.1. *Para todo $n \in \mathbb{N}$ e $j \in S$ tem-se,*

$$\pi_j = \sum_{i \in S} \pi_j \, p_{ij}^{(n)}.$$

Observações:

1) Na bibliografia é possível encontrar condições suficientes para a existência de distribuição limite de uma cadeia de Markov homogénea. Tais condições envolvem a classificação dos diferentes estados da cadeia e podem ser encontradas, por exemplo, em [35] ou em [44].

2) Este teorema garante que, se a cadeia de Markov possuir distribuição limite, então ela possuí distribuição estacionária e, para além disso, as duas distribuições coincidem. No entanto, o recíproco não é verdadeiro. Em Müller ([35], página 106-107), é fornecido um exemplo de uma cadeia de Markov que tem uma distribuição estacionária mas que não tem distribuição limite.

3.3 Cadeia de Markov de tempo contínuo

3.3.1 Definição e probabilidades de transição

Seja $\{X(t),\ t \geq 0\}$ uma família de variáveis aleatórias, assumindo valores num conjunto S, finito ou infinito numerável, e indexadas no conjunto $T = [0, +\infty[$. Se assumirmos que $S = \mathbb{N}_0$, sem perda de generalidade, diz-se que o processo $\{X(t),\ t \geq 0\}$ é uma cadeia de Markov de tempo contínuo se, para todos os instante $s, t \geq 0$ e para todos os inteiros não negativos $i, j, x(u),\ 0 \leq u < s$, se tem

$$P\left(X(t+s) = j | X(s) = i, X(u) = x(u),\ 0 \leq u < s\right) = P\left(X(t+s) = j | X(s) = i\right). \tag{3.10}$$

Se as probabilidades $P(X(t+s) = j | X(s) = i)$ não dependerem de s, então diz-se que a cadeia de Markov em tempo contínuo é homogénea. Neste caso, tais probabilidades denotam-se por apenas $p_{ij}(t)$, isto é,

$$p_{ij}(t) \equiv P\left(X(t+s) = j | X(s) = i\right).$$

Tal como aconteceu com as cadeias de Markov de tempo discreto, também aqui será dada mais atenção às cadeias homogéneas.

Se considerarmos uma cadeia de Markov homogénea $\{X(t), t \geq 0\}$ que começou no estado i, no instante 0, qual será a distribuição do tempo que a cadeia permanece nesse estado i, antes de mudar para um outro estado? Vamos representar esse tempo por T_i, com $i \in \mathbb{N}_0$, e mostrar que T_i tem a propriedade da falta de memória. De facto, para todo $s, t \geq 0$, tem-se:

$$\begin{aligned} P\left(T_i > t+s | T_i > s\right) &= \frac{P(T_i > t+s)}{P(T_i > s)} \\ &= \frac{P(X(t+s) = i \cap (X(u) = i, 0 \leq u \leq t+s))}{P(X(u) = i, 0 \leq u \leq s)} \\ &= P(X(t+s) = i \cap (X(u) = i, s < u < s+t) | X(u) = i, 0 \leq u \leq s) \\ &= P(X(u) = i, s < u \leq t+s | X(s) = i) \\ &= P(X(t) = i, 0 < u \leq t | X(0) = i) \\ &= P(T_i > t), \end{aligned}$$

tendo sido usada na quarta igualdade a propriedade Markoviana e na quinta igualdade a homogeneidade. Concluímos assim que T_i tem uma distribuição exponencial.

Exemplo 3.3.1. *Considere uma cadeia de Markov $\{X(t), t \geq 0\}$ de tempo contínuo homogénea que se inicia no estado i no instante 0. Sabendo que a cadeia não deixou o estado i (isto é, não ocorreu qualquer transição) durante os primeiros 10 minutos, qual é a probabilidade de a cadeia não sair do estado i durante os 5 minutos seguintes?*

***Solução**: Pretende-se calcular a seguinte probabilidade*

$$P(X(u) = i,\ 10 < u \leq 15 | X(u) = i,\ 0 \leq u \leq 10).$$

Uma vez que a variável T_i tem distribuição exponencial, com um certo parâmetro λ_i, a probabilidade pretendida é dada por

$$P\left(T_i > 15 | T_i > 10\right) = P\left(T_i > 5\right) = e^{-5\lambda_i},$$

tendo sido usado a propriedade da falta de memória da variável T_i na primeira igualdade.

O raciocínio efetuado para as variáveis T_i acima fornece, segundo ROSS [44], uma maneira simples de definirmos uma cadeia de Markov de tempo contínuo e homogénea. Podemos pensar numa tal cadeia como sendo um processo estocástico que se comporta do seguinte modo: quando entra no estado i,

i. o processo permanece nesse estado durante um tempo que tem distribuição exponencial com parâmetro que depende de i, digamos v_i;

ii. quando o processo sai do estado i, ele entra num outro estado $j \neq i$ com probabilidade P_{ij}, que satisfaz $\sum_{j \neq i} P_{ij} = 1$;

iii. a escolha do próximo estado a visitar após o estado i é independente do tempo gasto no estado i.

Observe-se que esta última condição tem que ser verificada para que a propriedade Markoviana seja mantida.

Apesar de poderem ocorrer outras situações, aqui vamos considerar sempre o caso em que $0 < v_i < \infty$ para todo estado i. Assim, podemos pensar numa cadeia de Markov de tempo contínuo como sendo um processo que se comporta como uma cadeia de Markov de tempo discreto, mas que é tal que o tempo que permanece em cada estado tem distribuição exponencial.

Observe-se ainda que, como v_i é a taxa de saída do estado i e P_{ij} é a probabilidade de o processo seguir do estado i para o estado j então, se consideramos a quantidade $v_i P_{ij}$ obtemos a taxa a que é feita a transição do estado i para o estado j, dado que a cadeia se inicia no estado i.

Vamos voltar a considerar as probabilidades de transição de uma cadeia de Markov de tempo contínuo e homogénea, isto é, as quantidades

$$p_{ij}(t) = P(X(t) = j | X(0) = i).$$

Neste tipo de processos, é usual considerarem-se as derivadas das respetivas probabilidades de transição, em ordem ao tempo, em $t = 0$. Tais derivadas serão designadas por intensidades de transição e estarão, obviamente, relacionadas com as quantidades v_i e P_{ij} atrás referidas.

Denote-se então por q_i a intensidade de passagem pelo estado i, dado que a cadeia se inicia no estado i. q_i é então dada pela seguinte derivada, quando existe:

$$q_i = -\frac{d}{dt} p_{ii}(0) = \lim_{t \to 0} \frac{p_{ii}(0) - p_{ii}(t)}{t} = \lim_{t \to 0} \frac{1 - p_{ii}(t)}{t}, \quad \text{para } i = 0, 1, ... \tag{3.11}$$

Denote-se ainda por q_{ij} a intensidade de transição do estado i para o estado j $(i \neq j)$, dado que a cadeia se inicia no estado i. Quando a seguinte derivada existir, q_{ij} é dada por

$$q_{ij} = \frac{d}{dt} p_{ij}(0) = \lim_{t \to 0} \frac{p_{ij}(t) - p_{ij}(0)}{t} = \lim_{t \to 0} \frac{p_{ij}(t)}{t}. \tag{3.12}$$

Trabalhando as equações (3.11) e (3.12) podemos escrever

$$\begin{cases} 1 - p_{ii}(t) = q_i t + o(t), \\ p_{ij}(t) = q_{ij} t + o(t), \text{ com } \frac{o(t)}{t} \underset{t \to 0}{\longrightarrow} 0 \end{cases}.$$

Estas igualdades mostram que as probabilidades de transição são assintoticamente proporcionais à amplitude do intervalo, isto é, a t, sendo que a primeira igualdade estabelece a proporcionalidade da transição do estado i para outro estado qualquer e a segunda igualdade estabelece a proporcionalidade da transição do estado i para um estado específico j $(j \neq i)$.

Observação: Em ROSS [44], estabelece-se as seguintes relações entre as quantidades v_i e P_{ij}, referidas anteriormente, e as quantidades q_i e q_{ij} acabadas de definir:

$$\begin{cases} q_i = v_i, \\ q_{ij} = v_i P_{ij}, \quad \text{para todos os estados } \ i, j \end{cases}.$$

3.3.2 Equações diferenciais de Kolmogorov

Analogamente ao que acontece no caso discreto, é fácil mostrar que estas probabilidades satisfazem as **equações de Chapman-Kolmogorov**, isto é, para todos os estados i e j e todos os instantes $h, t \geq 0$, tem-se

$$p_{ij}(t+h) = \sum_{k=0}^{\infty} p_{ik}(t) p_{kj}(h) \tag{3.13}$$

Se subtraimos $p_{ij}(t)$ em ambos membros da equação (3.13) obtém-se

$$\begin{aligned} p_{ij}(h+t) - p_{ij}(t) &= \sum_{k=0}^{\infty} p_{ik}(h) p_{kj}(t) - p_{ij}(t) \\ &= \sum_{k \neq i} p_{ik}(h) p_{kj}(t) - [1 - p_{ii}(h)] p_{ij}(t) \end{aligned}$$

e assim

$$\lim_{h \to 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} = \lim_{h \to 0} \left\{ \sum_{k \neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) - \left[\frac{1 - p_{ii}(h)}{h} \right] p_{ij}(t) \right\}.$$

Uma vez que, neste caso, se pode trocar o limite e a soma, se usarmos as equações (3.11) e (3.12), obtém-se as chamadas **equações diferenciais de Kolmogorov regressivas**, que estabelecem o seguinte: para todos os estados i, j e todo o instante $t \geq 0$,

$$p'_{ij}(t) = \sum_{k \neq i} q_{ik} p_{kj}(t) - q_j p_{ij}(t). \tag{3.14}$$

A partir da equação (3.13) podemos ainda também obter outras equações diferenciais. Subtraindo ambos os membros da equação (3.13) por $p_{ij}(t)$ obtém-se

$$\begin{aligned} p_{ij}(t+h) - p_{ij}(t) &= \sum_{k=0}^{\infty} p_{ik}(t) p_{kj}(h) - p_{ij}(t) \\ &= \sum_{k \neq j} p_{ik}(t) p_{kj}(h) - [1 - p_{jj}(h)] p_{ij}(t) \end{aligned}$$

e assim

$$\lim_{h \to 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} = \lim_{h \to 0} \left\{ \sum_{k \neq j} p_{ik}(t) \frac{p_{kj}(h)}{h} - \left[\frac{1 - p_{jj}(h)}{h} \right] p_{ij}(t) \right\}.$$

Assumindo que se pode trocar o limite e a soma, se usarmos as equações (3.11) e (3.12), obtém-se as chamadas **equações diferenciais de Kolmogorov progressivas**, que estabelecem o seguinte: para todos os estados i, j e todo o instante $t \geq 0$,

$$p'_{ij}(t) = \sum_{k \neq j} p_{ik}(t) q_{kj} - p_{ij}(t) q_j. \tag{3.15}$$

A troca do limite com a soma atrás referida nem sempre pode ser feita. No entanto, na maior parte dos modelos, incluindo as filas de espera Markovianas abordadas no capítulo seguinte, é possível fazer esta troca, conforme em ROSS [44].

3.3.3 Teoremas limite e distribuição estacionária

Tal como foi feito no caso discreto, vamos agora enunciar alguns resultados relativos ao comportamento limite de uma cadeia de Markov homogénea de tempo contínuo, sendo o limite estudado quando o tempo tende para o infinito, isto é, quando $t \to \infty$.

A definição de distribuição limite no caso em que o tempo é contínuo é semelhante à do tempo discreto, com uma evidente adaptação. Temos assim a seguinte definição:

Definição 3.3.1. *Seja* $\{X(t), t \geq 0\}$ *uma cadeia de Markov homogénea de tempo contínuo e* $(p_j, j \in \mathbb{N}_0)$ *uma distribuição de probabilidade sobre* $\mathbb{N}_0$. *Diz-se que* $(p_j, j \in \mathbb{N}_0)$ *é a distribuição limite da cadeia de Markov se*

$$\lim_{t\to\infty} p_{ij}(t) > 0 \quad e \quad \lim_{t\to\infty} p_{ij}(t) = p_j$$

para todos os estados i, j.

Observações:

1) Tal como acontece no caso discreto, se $(p_j, j \in \mathbb{N}_0)$ é a distribuição limite da cadeia de Markov então também é uma **distribuição estacionária**, isto é,

$$p_j = \sum_{i=0}^{\infty} p_i p_{ij}(t) \ \ \forall t > 0, \ \ \forall j \in \mathbb{N}_0.$$

2) Tal como aconteceu no caso discreto, existindo a distribuição limite $(p_j, j \in \mathbb{N}_0)$ este coincide com $\lim_{t\to\infty} P(X(t) = j)$. Isto demonstra-se facilmente. Assim,

$$\begin{aligned}
\lim_{t\to\infty} P(X(t) = j) &= \lim_{t\to\infty} P\left(X(t) = j \bigcap \left(\bigcup_{i\in\mathbb{N}_0} X(0) = i\right)\right) \\
&= \lim_{t\to\infty} \sum_{i\in\mathbb{N}_0} P(X(t) = j | X(0) = i) P(X(0) = i) \\
&= \sum_{i\in\mathbb{N}_0} \lim_{t\to\infty} P(X(t) = j | X(0) = i) P(X(0) = i) \\
&= p_j \sum_{i\in\mathbb{N}_0} P(X(0) = i) \\
&= p_j
\end{aligned}$$

tendo sido usado na segunda igualdade o teorema da probabilidade total e na quarta igualdade o facto de $(p_j, j \in \mathbb{N}_0)$ ser distribuição limite.

Do exposto nas observações, podemos constatar que a distribuição limite representa a probabilidade de a cadeia atingir um certo estado j ao fim de um longo período de tempo. Nestas condições dizemos que a cadeia alcançou um *estado de equilíbrio* ou um *estado estável*, sendo a distribuição de probabilidade limite $(p_j, j \in \mathbb{N}_0)$, independente do tempo e do estado inicial da cadeia.

Em condições gerais, a distribuição limite $(p_j, j \in \mathbb{N}_0)$ pode ser determinada à custa das equações diferenciais de Kolmogorov progressivas (3.15). Como, para todos os estados i, j

$$\lim_{t\to\infty} p_{ij}(t) = p_j \quad \text{e} \quad \lim_{t\to\infty} p'_{ij}(t) = 0, \tag{3.16}$$

então, tomando $\lim_{t\to\infty}$ em ambos os membros da equação

$$p'_{ij}(t) = \sum_{k\neq j} p_{ik}(t)q_{kj} - p_{ij}(t)q_j.$$

tem-se

$$p_j q_j = \sum_{k\neq j} p_k q_{kj}. \tag{3.17}$$

Conjugando a condição $\sum_{j=0}^{\infty} p_j = 1$ com a equação (3.17) é possível, regra geral, obter a distribuição limite $(p_j, j \in \mathbb{N}_0)$ de uma cadeia de Markov homogénea de tempo contínuo, quando tal distribuição existe.

Na secção a seguir, apresentamos o processo de contagem com especial atenção para o processo de Poisson. Estes processos serão usados nos modelos de filas de espera para descrever a chegada e saída de clientes ao sistema. Tais processos são de tempo contínuo, sendo o de Poisson uma cadeia de Markov.

3.4 Processo de contagem e processo de Poisson

3.4.1 Definições

Suponhamos que nos era pedido um modelo para contar, por exemplo, o número de ambulâncias que chegam ao serviço de urgência de um hospital, o número de chegadas de clientes a determinado serviço, o número de saídas de um produto num armazém, o número de chamadas telefónicas que chega a um posto de serviço, etc. Pode-se descrever estes tipos de fenómenos aleatórios através de uma função de contagem, denotada por $\{N(t),\ t \geq 0\}$, em que $N(t)$ representa o número de vezes que determinado acontecimento de interesse ocorre no intervalo de tempo $[0, t]$. A seguir apresentamos a definição formal de processos de contagem.

Definição 3.4.1. *Um processo estocástico de tempo contínuo $\{N(t),\ t \geq 0\}$ diz-se um processo de contagem se satisfaz as seguintes condições:*

i. $N(t) \geq 0$;

ii. o processo tem espaço de estado $\mathbb{N}_0$;

iii. se $s < t$, *então* $N(s) \leq N(t)$;

iv. para $s < t$, $N(t) - N(s)$ *representa o número de vezes que um determinado acontecimento ocorre no intervalo* $(s, t]$.

Um dos processos de contagem mais utilizados é o chamado processo de Poisson, que vamos definir em seguida.

Definição 3.4.2. *Um processo de contagem* $\{N(t),\ t \geq 0\}$ *é um processo de Poisson, com intensidade (ou taxa)* $\lambda > 0$*, se satisfaz os seguintes três axiomas:*

A1. $N(0) = 0$*, isto é, no instante inicial nenhum acontecimento ocorreu;*

A2. O processo tem incrementos independentes, isto é, para quaisquer instantes $0 = t_0 < t_1 < t_2 < ... < t_{n-1} < t_n$ *as* n *variáveis aleatórias*

$$N(t_1) - N(t_0), N(t_2) - N(t_1), ..., N(t_n) - N(t_{n-1})$$

são independentes;

A3. O número de acontecimentos em qualquer intervalo de amplitude t *tem distribuição de Poisson com média* λt*, isto é, para todo* $s, t \geq 0$

$$P(N(t+s) - N(s) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!},\ n = 0, 1, 2, ...$$

Obsevações:

1) Segue do axioma $A3$ que um processo de Poisson é tal que

$$E[N(t)] = \lambda t.$$

Isto justifica a designação de taxa para o parâmetro λ. Note-se ainda que λ representa o número médio de acontecimentos por unidade de tempo.

2) Do axioma $A3$ podemos concluir ainda que, para todo o $h > 0$, as variáveis aleatórias

$$N(t_2 + h) - N(t_1 + h) \ \text{ e } \ N(t_2) - N(t_1),$$

com $t_1, t_2 \geq 0$, têm a mesma distribuição. Isto significa que o processo de Poisson tem incrementos estacionários. Deste modo, um processo de Poissson tem incrementos independentes (axioma $A2$) e estacionários

Os axiomas $A1$ e $A2$ desta definição de processo de Poisson são, em geral, facilmente verificado. No entanto, o axioma $A3$ é, regra geral, mais difícil de verificar e, portanto, pode ser útil ter uma definição alternativa. Vamos então enunciar uma outra definição de processo de Poisson equivalente à anterior.

Definição 3.4.3. *Um processo de contagem* $\{N(t),\ t \geq 0\}$ *é um processo de Poisson, com intensidade (ou taxa)* $\lambda > 0$*, se satisfaz os seguintes quatro axiomas:*

*A1** $N(0) = 0$*;*

A2 o processo tem incrementos independentes e estacionários;*

*A3** $P(N(h) = 1) = \lambda h + o(h)$*;*

*A4** $P(N(h) \geq 2) = o(h)$.

Demonstrar que esta definição implica a anterior não é complicado e pode ser consultado em ROSS ([44], pag. 315). Aqui vamos demonstrar que a definição anterior implica esta última. Os axiomas $A1^*$ e $A2^*$ são consequências diretas do axiomas $A1$, $A2$ e $A3$. Para demonstrar os axiomas $A3^*$ e $A4^*$, deve-se recorrer a seguinte expansão:

$$e^x = \sum_{k=0}^{+\infty} \frac{x^k}{k!},\ x \in \mathbb{R}.$$

Deste modo, usando o axioma $A3$,

- demonstra-se $A3^*$,

$$P(N(t) = 1) = e^{-\lambda h}\lambda h = \lambda h \left[1 - \lambda h + \frac{(\lambda h)^2}{2!} + ...\right] = \lambda h + o(h);$$

- demonstra-se o $A4^*$,

$$P(N(h) \geq 2) = e^{-\lambda h}\left[\frac{(\lambda h)^2}{2!} + \frac{(\lambda h)^3}{3!} + ...\right] = o(h).$$

3.5 Tempos de espera e tempos entre chegadas

Seja $\{N(t), t \geq 0\}$ um processo de contagem e sejam $\tau_1, \tau_2, ...$ $(0 < \tau_1 < \tau_2 < ...)$ os instantes no tempo em que ocorre o acontecimento de interesse. Assim, τ_i representa o tempo de espera até a i-ésima ocorrência do acontecimento de interesse, $i = 1, 2, ...$

A sucessão de variáveis aleatórias $\{T_n, n \geq 1\}$ definidas por

$$T_1 = \tau_1, T_2 = \tau_2 - \tau_1, ..., T_n = \tau_n - \tau_{n-1}, ... \tag{3.18}$$

vai representar a sucessão dos tempo entre duas ocorrências sucessivas do acontecimento. Isto é, T_n vai representar o tempo decorrido entre a $(n-1)$-ésima ocorrência e a n-ésima ocorrência.

Em determinados contextos, em particular numa fila de espera, $\{\tau_n, n \geq 1\}$ e $\{T_n, n \geq 1\}$ são conhecidas como a sucessão dos **tempos de espera** e sucessão dos **tempos entre chegadas**, respetivamente. Observe que, a equacao (3.18) permite concluir que

$$\tau_1 = T_1, \tau_2 = T_1 + T_2, ..., \tau_n = T_1 + ... + T_n, ...$$

Segundo Müller [35], um processo de contagem $\{N(t), t \geq 0\}$ e a respetiva sucessão dos tempos de espera $\{\tau_n, n \geq 1\}$ satisfazem as seguintes relações de equivalência:

a) Para todo $t > 0$ e $n = 1, 2, ...,$

$$N(t) \leq n \Leftrightarrow \tau_{n+1} > t.$$

Deste modo, afirmar que o número de acontecimentos que ocorrem em $[0,t]$ é menor ou igual a n é equivalente a dizer que o período de tempo até que ocorra o $(n+1)$-ésimo acontecimento é superior a t.

b) Para todo $t>0$,

$$N(t)=0 \Leftrightarrow \tau_1 > t.$$

De facto, se até ao instante t não ocorreu nenhum acontecimento é equivalente a dizer-se que o acontecimento ocorrerá pela primeira vez num instante superior a t.

c) Para todo $t>0$ e $n=1,2,...,$

$$N(t)=n \Leftrightarrow \tau_n \leq t \text{ e } \tau_{n+1} > t.$$

Isto significa, que a ocorrência de exatamente n acontecimento em $[0,t]$ é equivalente a que o n-ésimo acontecimento ocorra num instante inferior ou igual a t e que o $(n+1)$-ésimo acontecimento ocorra depois de t.

Estas relações têm as seguintes implicações em termos probabilísticos:

i) $F_{N(t)}(n) = P(N(t) \leq n) = 1 - F_{\tau_{n+1}}(t), n=1,2,...;$

ii) $P(N(t)=0) = 1 - F_{\tau_1}(t);$

iii)

$$\begin{aligned} P(N(t)=n) &= P(\tau_n \leq t, \tau_{n+1} > t) = P((\tau_n \leq t) \cap \overline{(\tau_{n+1} \leq t)}) \\ &= P(\tau_n \leq t) - P(\tau_{n+1} \leq t) = F_{\tau_n}(t) - F_{\tau_{n+1}}(t). \end{aligned}$$

Vamos agora enunciar alguns resultados relativos à distribuição dos tempos de espera e dos tempos entre chegadas num processo de Poisson.

Teorema 3.5.1. *Seja $\{N(t), t \geq 0\}$ um processo de Poisson de intensidade (ou taxa) $\lambda > 0$. Então, para cada n, τ_n tem distribuição Gama de parâmetros n e λ, ou seja, a sua função densidade de probabilidade é,*

$$f_{\tau_n}(t) = \begin{cases} \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}, & se \quad t \geq 0 \\ 0, & se \quad t < 0 \end{cases}.$$

Para a demonstração deste teorema, observe que, para todo $t>0$,

$$\begin{aligned} F_{\tau_n}(t) &= P(\tau_n \leq t) = 1 - P(\tau_n > t) \\ &= 1 - P(N(t) \leq n-1) \\ &= 1 - \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!} \end{aligned}$$

Derivando em ordem a t, obtém-se a função densidade de probabilidade acima mencionada. De facto,

$$F'_{\tau_n}(t) = e^{-\lambda t}\left[\sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!} - \sum_{k=1}^{n-1}\frac{(\lambda t)^{k-1}}{(k-1)!}\right] = \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}.$$

Relativamente à sucessão dos tempos entre chegadas, temos os seguintes teoremas.

Teorema 3.5.2. *Seja $\{N(t), t \geq 0\}$ um processo de Poisson de intensidade (ou taxa) $\lambda > 0$. Então a sucessão das variáveis aleatórias dos tempos entre chegadas, $\{T_n, n \geq 1\}$ é independente e identicamente distribuída com a distribuição exponencial de valor médio $\frac{1}{\lambda}$.*

Teorema 3.5.3. *Se sucessão dos tempos entre chegadas $\{T_n, n \geq 1\}$ é independente e identicamente distribuída com a distribuição exponencial de valor médio $\frac{1}{\lambda}$, então o correspondente processo de contagem $\{N(t), t \geq 0\}$ é de Poisson com intensidade (ou taxa) λ.*

Exemplo 3.5.1. *Considere que as pessoas imigram para um país de acordo com um processo de Poisson com a intensidade (ou taxa) $\lambda = 2$ por dia. Calcule:*

a) O tempo esperado até que o décimo imigrante chegue neste pais?

b) A probabilidade de que o tempo decorrido entre a décima e a décima primeira chegada exceda dois dias?

***Solução**:*

a) Seja T_1 o tempo de chegada do primeiro imigrante e T_n o tempo entre a chegada do $(n-1)$-ésimo e do n-enésimo imigrante $(2 \leq n \leq 10)$. Então, $T_1, T_2, ..., T_{10}$ são variáveis aleatórias exponenciais, independente identicamente distribuídas, com media $\frac{1}{\lambda}$. Assim, estamos interessado em calcular $E[\tau_{10}]$.

$$E[\tau_{10}] = E\left[\sum_{n=1}^{10} T_n\right] = \frac{10}{2} = 5$$

logo, o tempo esperado até que o décimo imigrante chegue nestes pais é de 5 dias.

b) Seja $T_{11} - T_{10}$ o tempo decorrido entre a décima e a décima primeira chegada. Usando a propriedade falta de memoria (equação (2.2)) temos:

$$P(T_{11} - T_{10} > 2) = e^{-4} \cong 0.0183$$

logo, a probabilidade de que o tempo decorrido entre a décima e a décima primeira chegada exceda dois dias é de 0.0183.

Capítulo 4

Teoria de Filas de Espera

4.1 Introdução

A teoria de filas de espera fornece modelos probabilísticos que permitem estudar a formação de filas em função do número de chegadas e de processo de atendimento de "clientes". Usando ferramentas matemáticas, esta teoria permite obter informação importante sobre diferentes aspetos da fila de espera e, em particular, permite encontrar situações de equilíbrio que satisfaçam o cliente e sejam viáveis ao servidor.

Segundo Müller [35], uma situação de fila de espera é caracterizada por um fluxo de clientes ou utentes que chegam a um ou mais postos de serviço (por exemplo, supermercados, bancos, correios, portagem de auto-estrada,....) a fim de satisfazer uma qualquer necessidade. Quando o número de clientes é maior do que número de postos de serviço, então forma-se o que usualmente se chama de uma fila de espera.

Numa situação de fila de espera estão presentes duas principais características aleatórias muito importantes. A primeira, diz respeito ao processo de chegada dos clientes que decorre de forma aleatória ao longo do tempo, e a segunda, ao tempo de serviço correspondente à ocupação do respetivo posto de serviço por cada cliente.

Um sistema de fila de espera considera-se como o conjunto constituído pela fila de espera, propriamente dita, e pelos respetivos postos de serviço. Num dado instante, o estado do sistema é dado pela soma do número de clientes que aguardam na fila com o número de clientes que estão a ser servidos nesse mesmo instante (isto é, que estão a ocupar um posto de serviço).

Nos modelos de filas de espera que vamos estudar, estaremos interessados em estudar, entre outras coisas, caraterísticas como o número médio de clientes no sistema (ou apenas o número de clientes que aguardam na fila) e o tempo médio que um cliente gasta no sistema (ou apenas o tempo que um cliente gasta a aguardar na fila).

4.2 Estrutura de um sistema de filas de espera

A estrutura de um sistema de fila de espera tem em conta vários elementos, como por exemplo: a população ou fonte, a fila, o serviço, capacidade máximo de clientes no sistema e a disciplina do serviço.

4.2.1 População ou fonte

A população desempenha um papel relevante no sistema de filas de espera, sendo a fonte de clientes que se dirigem ao sistema. Pode ser considerada finita ou infinita (de modo que a fonte de entrada também é dita limitada ou ilimitada).

Distribuição das chegadas ao sistema
As chegadas podem ser descritas pelo tempo que decorre entre duas chegadas sucessivas (tempo entre chegadas) ou pelo número de chegadas por unidade de tempo (distribuição das chegadas). Podemos ter chegadas:

- constantes - quando os intervalos de tempo entre chegadas sucessivas são fixos;
- aleatórias - quando os intervalos de tempos entre chegadas sucessivas não podem ser previstos com certeza, pelo que neste caso usam-se distribuições de probabilidade para modelar os tempos entre chegadas.

Taxa de chegada
Corresponde ao número médio de clientes que chegam ao sistema por unidade de tempo. Pode ser:

- independente do estado do sistema (é usualmente denotada por λ);
- dependente do estado do sistema (é usualmente denotada por λ_n, onde n é o número de clientes no sistema).

Comportamento dos clientes
Distinguem-se essencialmente dois tipos de comportamentos:

- paciente - quando os clientes ficam na fila de espera até serem servidos;
- impaciente - quando os clientes desistem de esperar depois de terem estado algum tempo na fila ou quando existem clientes que se recusam a juntar à fila de espera por esta ser demasiada longa.

4.2.2 Fila de espera

A fila de espera é onde os clientes esperam antes de ser servidos. São classificadas em função do seu número e seu comprimento.

Número de filas de espera

- filas de espera simples - são consideradas filas de espera únicas, mesmo que o sistema tenha vários postos de serviço;
- filas de espera múltiplas - são consideradas filas de espera por cada posto de serviço.

Comprimento da fila de espera
O comprimento de uma fila de espera pode ser finito ou infinito.

4.2.3 Serviço

Configuração do serviço
A configuração do serviço consiste, regra geral, no número de postos de serviços em paralelo. Podem existir situações em que na configuração de serviço são desenhadas varias fases de atendimento. Tal acontece por exemplo, quando um cliente tem que ser atendido por vários balcões numa mesma entidade ou quando o serviço opera de maneiras diferentes ao longo do tempo (pode depender da altura do dia, da altura do ano ou até do estado do tempo).

Dimensão do serviço
Podemos ter serviço:

- simples - cada posto atende um cliente de cada vez (como por exemplo num supermercado);
- em grupo - cada posto pode atender vários clientes em simultâneo (como por exemplo num elevador).

Distribuição dos tempos de serviço
A distribuição do tempo de serviço pode ser constante ou aleatória. Quando é aleatória, as distribuições mais frequentemente adotadas são a distribuição Exponencial e a Gama.

Taxa de serviço
Corresponde ao número médio de clientes atendidos por unidade de tempo. Pode ser:

- independente do estado do sistema (é usualmente denotada por μ);
- dependente do estado do sistema (é usualmente denotada por μ_n, onde n é o número de clientes no sistema).

Observe-se que a taxa de serviço, μ ou μ_n, é o número médio de clientes servidos por posto e por unidade de tempo, a que corresponde um tempo médio de serviço por posto igual a $\frac{1}{\mu}$ ou $\frac{1}{\mu_n}$ respetivamente.

4.2.4 Capacidade máxima do sistema

A capacidade máxima do sistema corresponde ao número máximo de clientes que o sistema pode suportar. Esta capacidade pode ser finita ou infinita.

4.2.5 Disciplina de serviço

A disciplina de serviço estabelece a maneira como se selecionam os clientes da fila de espera para o posto de serviço. Existem várias disciplinas de serviço:

- FIFO **"first in first out"** ou FCFS **"fisrt come first served"** - este código para disciplina de serviço é o mais comum e é utilizado em filas de espera onde os clientes são servidos por ordem de chegada;
- LIFO **"last in first out"** ou LCFS **"last come first served"** - este código para disciplina de serviço é usado em filas de espera onde o último cliente a chegar é o primeiro a ser servido e a sair do sistema;
- SIRO **"service in random order"** - este código para disciplina de serviço é usado em filas de espera onde o serviço é feito de forma aleatória;
- GD **"general discipline"** - este código para disciplina de serviço é usado em filas de espera onde não se especifica a disciplina de serviço.

4.3 Característica de uma fila de espera

Segundo Müller [35], a descrição das características de um sistema de filas de espera foi sugerida pelo matemático inglês David Geoge Kendall (1918-2007). Essa descrição tem a seguinte notação:

$$(a/b/c) : (d/e/f)$$

onde:

- a é a distribuição dos tempos entre chegadas sucessivas de clientes;
- b é a distribuição dos tempos de serviço;
- c é o número de postos de serviço, onde $c \in \mathbb{N}$;
- d é a disciplina de serviço;
- e é a capacidade do sistema, onde $e \in \mathbb{N} \cup \{+\infty\}$;
- f é o tamanho da população, onde $f \in \mathbb{N} \cup \{+\infty\}$.

Para as notações a e b de uma fila de espera podemos ter as seguintes possibilidades:

- M = distribuição exponencial;

- D = tempos determinísticos;
- E = distribuição Gama;
- G = distribuição não é especificada.

Existem outras possibilidades para a e b, mais não serão abordada neste trabalho.

Em muitos modelos de filas de espera, as indicações de d, e e f são omitidas. Tal significa que se assume que a disciplina de serviço é FIFO, a capacidade do sistema é ilimitado e a população é infinita.

4.4 Terminologia e notação de uma fila de espera

Para as filas de espera será utilizada a seguinte terminologia e notação padrão:

- estado do sistema= número de clientes que estão a ser servidos somado com o número de clientes que estão à espera;
- comprimento da fila = número de clientes à espera de serviço, isto é, estado do sistema subtraído do número de clientes que estão a ser servidos;
- $N(t)$ = variável aleatória que representa o estado do sistema no instante t ($t \geq 0$);
- $p_n(t) = P(N(t) = n)$;
- p_n = probabilidade de existirem n clientes no sistema quando este começou a operar há bastante tempo, ou seja, quando o sistema está em equilíbrio;
- λ = intensidade (ou taxa) de chegada de clientes por unidade de tempo;
- $\frac{1}{\lambda}$ = tempo médio entre chegadas sucessivas de clientes;
- λ_n = taxa de chegada de novos clientes quando n clientes já estão no sistema;
- $\overline{\lambda}$ = taxa média de chegada de clientes no sistema.

$$\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n p_n; \tag{4.1}$$

- μ = taxa de serviço de cada um dos postos de serviço, ou seja, número esperado de clientes servidos por unidade de tempo em cada posto de serviço;
- $\frac{1}{\mu}$ = tempo médio de serviço de cada um dos postos de serviço;
- μ_n = taxa de serviço quando n clientes estão no sistema;
- s = sistema;
- q = fila;

- L_s = número médio de clientes no sistema;
- L_q = número médio de clientes que aguardam na fila, isto é, comprimento médio da fila;
- W_s = tempo médio de espera no sistema;
- W_q = tempo médio de espera na fila;
- ρ = taxa de ocupação do serviço (ou intensidade do tráfego), ou seja, percentagem esperada de tempo que um posto de serviço está ocupado. Em geral, é igual a $\frac{\lambda}{\mu}$;
- $1 - \rho$ = taxa de desocupação do serviço, ou seja, percentagem esperada de tempo que um posto de serviço está desocupado.

4.5 Estado estacionário ou de equilíbrio

Determinada notação também é necessária para descrever o estado estacionário ou de equilíbrio da fila de espera. Se um sistema de fila de espera tiver começado recentemente a operação, o estado do sistema será afetado pelo estado inicial e pelo tempo decorrido. Neste caso, o sistema é dito estar numa condição transitória. No entanto, depois de decorrido o tempo suficiente, o estado do sistema torna-se essencialmente independente do estado inicial e do tempo decorrido (excepto em circunstâncias incomuns) e o sistema é dito estar numa condição de equilíbrio.

Numa situação de equilíbrio do sistema de fila de espera, o número médio de clientes no sistema (L_s) e o número médio de clientes que aguardam na fila (L_q) definem-se por:

$$L_s = \sum_{n=0}^{\infty} n p_n \tag{4.2}$$

e

$$L_q = \sum_{n=c+1}^{\infty} (n-c) p_n, \tag{4.3}$$

onde c é o número de postos de serviços em paralelo e $\{p_n, n \in \mathbb{N}_0\}$ é a distribuição de equilíbrio do sistema.

John D.C.Little forneceu a primeira prova rigorosa de que, no estado estacionário ou de equilíbrio, a seguinte equação é valida:

$$L = \lambda W, \tag{4.4}$$

para o caso em λ, a taxa de chegada de clientes no sistema, é constante. Esta equação às vezes é referida como a fórmula de Little e permite-nos concluir que

$$L_s = \lambda W_s \tag{4.5}$$

e

$$L_q = \lambda W_q. \tag{4.6}$$

Se a taxa de clientes não for constante, então λ pode ser substituído nestas equações por $\overline{\lambda}$ dado em (4.1).

Quando o tempo médio de serviço de cada posto é $\frac{1}{\mu}$, então tem-se

$$W_s = W_q + \frac{1}{\mu}. \tag{4.7}$$

Conjugando esta equação e a equação (4.5) obtém-se

$$L_s = L_q + \rho, \tag{4.8}$$

em que $\rho = \frac{\lambda}{\mu}$.

4.6 Modelos de filas de espera

Nesta secção serão apresentados alguns dos modelos de fila de espera mais utilizados na prática. Especial destaque será dado aos modelos markovianos, mas serão apresentados também alguns não markovianos.

Os modelos markovianos são aqueles em que se assume que os clientes chegam ao sistema de acordo com um processo de Poisson de intensidade (ou taxa) λ, ou seja, a sucessão dos tempos entre chegadas é independente e identicamente distribuída com distribuição exponencial de parâmetro λ. Assume-se também que os tempos de serviço são independentes e identicamente distribuídos com distribuição exponencial de parâmetro μ.

Os modelos não-markovianos são aqueles em que as chegadas dos clientes e/ou os tempos de serviço seguem uma distribuição geral.

4.6.1 Modelo $(M/M/1) : (FIFO/\infty/\infty)$

Este modelo de filas de espera assume que o processo de chegada de clientes é um processo de Poisson, ou seja, os tempos entre chegadas sucessivas são independentes e identicamente distribuídos com uma distribuição exponencial. Assume também que os tempos de serviço são independentes e identicamente distribuídos com uma outra distribuição exponencial. Mais, o sistema tem um único servidor, a disciplina de serviço é $FIFO$ (primeiro cliente a chegar primeiro a ser servido e a sair), e a capacidade do sistema e o tamanho da população são infinitos. Um exemplo de aplicação deste modelo pode-se ver em [50], onde os autores usam o modelo para determinar o número adequado de pessoal prestador de serviço durante o turno da noite numa sala de cirurgia.

O modelo $(M/M/1) : (FIFO/\infty/\infty)$ pode ser representado através da Figura 4.1.

Temos assim um modelo em que a taxa de chegada de clientes é sempre igual a λ, isto é,

$$\lambda_n = \lambda, \text{ para } n = 0, 1, 2, ...$$

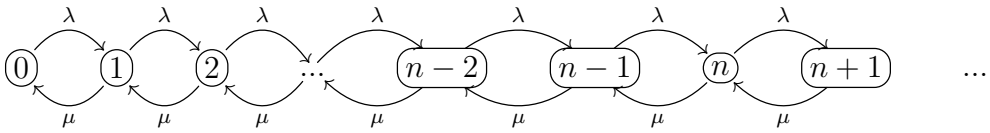


Figura 4.1: Modelo $(M/M/1) : (FIFO/\infty/\infty)$.

e a taxa de saída de clientes é sempre igual a μ, isto é,

$$\mu_n = \mu \text{ para } n = 1, 2, ...$$

A partir das equações diferenciais de Kolmogorov (3.17) vamos determinar a distribuição de equilíbrio (ou estacionária) $(p_k, \ \ k \in \mathbb{N}_0)$ para o modelo de filas de espera $(M/M/1) : (FIFO/\infty/\infty)$. Passamos a descrever as equações que vão então permitir determinar a distribuição estacionária para cada estado.

Estado 0
Usando a equação (3.17) com $j = 0$, obtemos:

$$p_0 q_0 = \sum_{k \neq 0} p_k q_{k0} \Leftrightarrow p_0 q_0 = p_1 q_{10} \Leftrightarrow p_0 \lambda = p_1 \mu \Leftrightarrow p_1 = \frac{\lambda}{\mu} p_0, \tag{4.9}$$

uma vez que $q_{k0} = 0, \ \text{ se } \ k \neq 1$.

Esta equação traduz a ideia de equilíbrio para o estado 0, ou seja, o fluxo de entradas deve ser igual ao fluxo de saídas. Assim, por palavras, podemos descrever esta equação do seguinte modo: a entrada no estado 0 ocorre quando o processo sai do estado 1 para o estado 0 devido a saída de um cliente do sistema (o fluxo de entradas no estado 0 é assim quantificado por μp_1) e a saída do estado 0 ocorre quando o processo sai do estado 0 para o estado 1 devido a chegada de um cliente à fila (o fluxo de saídas do estado 0 é assim quantificado por λp_0).

Estado 1
Usando a equação (3.17) com $j = 1$, obtemos:

$$p_1 q_1 = \sum_{k \neq 1} p_k q_{k1} \Leftrightarrow p_1 q_1 = p_0 q_{01} + p_2 q_{21} \Leftrightarrow p_1(\lambda + \mu) = p_0 \lambda + p_2 \mu \Leftrightarrow p_2 = \frac{\lambda}{\mu} p_1 + p_1 - \frac{\lambda}{\mu} p_0,$$

onde $p_1 = \frac{\lambda}{\mu} p_0$, pelo que

$$p_2 = \left(\frac{\lambda}{\mu}\right)^2 p_0. \tag{4.10}$$

Note-se que $q_{k1} = 0 \ \ \forall k \neq 0, 2$ e que $q_1 = \lambda + \mu$. Observe que $q_1 = \lambda + \mu$ é o parâmetro de uma $\text{Exp}(\lambda + \mu)$ uma vez que o sistema permanece no estado 1 durante um tempo aleatório que corresponde ao mínimo entre dois tempos exponenciais independentes, isto é, entre $\text{Exp}(\lambda)$ e $\text{Exp}(\mu)$.

Portanto, a entrada no estado 1 ocorre quando o processo sai do estado 0 para o estado 1 devido a chegada de um cliente a fila ou quando o processo sai do estado 2 para o estado 1 devido a saída de um cliente do sistema (o fluxo de entradas no estado 1 é assim quantificado por $\lambda p_0 + \mu p_2$) e a saída do estado 1 ocorre quando o processo sai do estado 1 para o estado 2 por chegada de um cliente à fila ou quando o processo sai do estado 1 para o esta 0 por saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $\lambda p_1 + \mu p_1$).

Estado 2

Usando a equação (3.17) com $j = 2$, obtemos:

$$p_2 q_2 = \sum_{k \neq 2} p_k q_{k2} \Leftrightarrow p_2 q_2 = p_1 q_{12} + p_3 q_{32} \Leftrightarrow p_2(\lambda + \mu) = p_1 \lambda + p_3 \mu \Leftrightarrow p_3 = \frac{\lambda p_2 + \mu p_2 - \lambda p_1}{\mu},$$

substituído os valores das equações (4.9) e (4.10) obtemos,

$$p_3 = \left(\frac{\lambda}{\mu}\right)^3 p_0. \tag{4.11}$$

Note-se que $q_{k2} = 0$ se $\forall k \neq 1, 3$.

A entrada no estado 2 ocorre quando o processo sai do estado 1 para o estado 2 devido a chegada de um cliente a fila ou quando o processo sai do estado 3 para o estado 2 devido a saída de um cliente do sistema (o fluxo de entradas no estado 2 é assim quantificado por $\lambda p_1 + \mu p_3$) e a saída do estado 2 ocorre quando o processo sai do estado 2 para o estado 3 devido a chegada de um cliente à fila ou quando o processo sai do estado 2 para o estado 1 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $\lambda p_2 + \mu p_2$).

Estado n

No estado n o raciocínio é o mesmo aos estados estudados anteriormente, e podemos concluir que,

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \Leftrightarrow p_n = \rho^n p_0. \tag{4.12}$$

em que $\rho = \frac{\lambda}{\mu}$ (taxa de ocupação).

Uma vez obtidos p_n, $n \geq 1$, p_0 é obtido a partir da seguinte equação:

$$\sum_{n=0}^{\infty} p_n = 1.$$

Então,

$$\sum_{n=0}^{\infty} p_n = 1 \Leftrightarrow \sum_{n=0}^{\infty} \rho^n p_0 \Leftrightarrow p_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1},$$

onde o somatório é a soma de todos termos de uma progressão geométrica de razão ρ ($\rho < 1$). Daí,

$$p_0 = \left(\frac{1}{1-\rho}\right)^{-1} \Leftrightarrow p_0 = 1 - \rho \tag{4.13}$$

e

$$p_n = (1-\rho)\rho^n, \text{ para } n = 0, 1, 2, ... \tag{4.14}$$

Estamos agora em condições de determinar o número médio de clientes no sistema, isto é,

$$L_s = \sum_{n=0}^{\infty} n p_n.$$

Usando a equação (4.14) obtemos:

$$\begin{aligned} L_s &= \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\sum_{n=1}^{\infty} n\rho^n = (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1} \\ &= (1-\rho)\rho\sum_{n=1}^{\infty}\frac{d}{d\rho}(\rho^n) = (1-\rho)\rho\frac{d}{d\rho}\left(\sum_{n=1}^{\infty}\rho^n\right) \\ &= (1-\rho)\rho\frac{d}{d\rho}\left(\frac{\rho}{1-\rho}\right) = \frac{\rho}{1-\rho}, \end{aligned}$$

Uma vez que $\rho = \frac{\lambda}{\mu}$ também se pode escrever

$$L_s = \frac{\lambda}{\mu - \lambda}, \tag{4.15}$$

Usando a equação (4.15), juntamente com as equações (4.8), (4.3) e (4.2), podemos determinar:

- o valor de L_q,

$$L_q = L_s - \rho = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{(\mu-\lambda)\mu} \qquad \text{ou} \qquad L_q = \frac{\rho^2}{1-\rho}.$$

- o valor de W_q,

$$W_q = \frac{L_q}{\lambda} = \frac{\frac{\lambda^2}{(\mu-\lambda)\mu}}{\lambda} = \frac{\lambda}{(\mu-\lambda)\mu} \qquad \text{ou} \qquad W_q = \frac{\rho}{\mu(1-\rho)}.$$

- o valor W_s,

$$W_s = \frac{L_s}{\lambda} = \frac{\frac{\lambda}{\mu-\lambda}}{\lambda} = \frac{1}{\mu-\lambda} \qquad \text{ou} \qquad W_s = \frac{1}{\mu(1-\rho)}.$$

Neste modelo de de filas de espera, podemos derivar a distribuição de probabilidade do tempo gasto no sistema por um cliente escolhido ao acaso. Vamos denotar esse tempo u_s e observar que incluí o tempo de serviço deste cliente. Se este cliente encontrar n clientes no sistema, então ele terá de esperar um tempo que corresponde a soma de $n+1$ tempos, incluindo o tempo do seu serviço, ou seja,

$$u_n = U_1 + U_2 + ... + U_n + U_{n+1},$$

onde U_1 é o tempo necessário para o cliente que está em serviço poder terminar, $U_2, ..., U_n$ são os tempos de serviço dos restantes $n-1$ clientes e U_{n+1} o tempo de serviço do cliente que acaba de chegar. Uma vez que $U_i, i = 1, ..., n+1$ são variáveis aleatórias independentes e identicamente distribuídas, com a distribuição exponencial de parâmetro μ, então u_n tem distribuição Gama com parâmetros $(\mu, n+1)$, ou seja,

$$P[u_s \leq t | n \text{ clientes no sistema}] = \int_0^t \frac{(\mu x)^n}{n!} \mu e^{-\mu x} dx, \ t \geq 0.$$

Aplicando o teorema da probabilidade total obtemos,

$$\begin{aligned}
P[u_s \leq t] &= \sum_{n=0}^{\infty} P[u_s \leq t | n \text{ cliente no sistema}].p_n \\
&= \sum_{n=0}^{\infty} \left[\int_0^t \frac{(\mu x)^n}{n!} \mu e^{-\mu x}.\rho^n(1-\rho) dx \right] \\
&= \sum_{n=0}^{\infty} \left[\int_0^t \frac{(\mu x)^n}{n!} \mu e^{-\mu x}.\left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) dx \right] \\
&= (\mu - \lambda) \int_0^t e^{-\mu x} \sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} dx = (\mu - \lambda) \int_0^t e^{-\mu x}.e^{\lambda x} dx \\
&= (\mu - \lambda) \int_0^t e^{-(\mu-\lambda)x} dx = 1 - e^{-(\mu-\lambda)t},
\end{aligned}$$

ou seja, u segue uma distribuição exponencial de parâmetro $(\mu - \lambda)$. Assim, a probabilidade de o tempo gasto no sistema ser maior que t é,

$$P(u_s > t) = 1 - P(u_s \leq t) = e^{-(\mu-\lambda)t}, \ \ t \geq 0.$$

A probabilidade de o tempo que um cliente aguarda na fila ser maior que t é dada por

$$P(u_q > t) = 1 - P(u_q \leq t)$$

onde

$$\begin{aligned} P(u_q \leq t) &= p_0 + \sum_{n=1}^{\infty} P[u_q \leq t | n \text{ cliente no sistema}].p_n \\ &= p_0 + \sum_{n=1}^{\infty} \left[\int_0^t \frac{(\mu x)^{n-1}}{(n-1)!} \mu e^{-\mu x}.\rho^n(1-\rho)dx \right] \\ &= p_0 + \sum_{n=1}^{\infty} \left[\int_0^t \frac{(\mu x)^{n-1}}{(n-1)!} e^{-\mu x}.\rho(\mu-\lambda)dx \right] = p_0 + \rho(\mu-\lambda) \int_0^t e^{-\mu x} \left(\sum_{n=1}^{\infty} \frac{(\mu x)^{n-1}}{(n-1)!} \right) dx \\ &= p_0 + \rho(\mu-\lambda) \int_0^t e^{-\mu x} e^{\lambda x} dx = \rho \left(1 - e^{-(\mu-\lambda)}\right), \end{aligned}$$

pelo que

$$P(u_q > t) = \rho P(u_s > t) = \frac{\lambda}{\mu} e^{-(\mu-\lambda)t}, \; para \; t \geq 0. \tag{4.16}$$

Assim, a probabilidade de um cliente não ter que esperar na fila é

$$P(u_q = 0) = 1 - P(u_q > 0) = 1 - \rho = p_0. \tag{4.17}$$

Seja agora N o número de cliente no sistema numa situação de equilíbrio. A probabilidade de existirem k ou mais clientes no sistema será dada por:

$$P(N \geq k) = \sum_{n=k}^{\infty} p_n = \sum_{n=k}^{\infty} \rho^n(1-\rho) = (1-\rho) \sum_{n=0}^{\infty} \rho^k \rho^n = (1-\rho)\rho^k \sum_{n=0}^{\infty} \rho^n = \rho^k.$$

Tabela 4.1: Tabela das características do modelo $(M/M/1) : (FIFO/\infty/\infty)$

Chegada : Poissoneana	Tempo de atendimento: exponencial
Taxa : λ clientes/unidade de tempo	Taxa : μ clientes/unidade de tempo
População = ∞	Nº de servidores = 1
Nº máximo de clientes na fila = ∞	Taxa de ocupação $\rho = \frac{\lambda}{\mu}$ com $\rho < 1$
	Taxa de desocupação = $1 - \rho$
Número médio de clientes no sistema	$L_s = \frac{\lambda}{\mu - \lambda}$
Número médio de clientes que aguardam na fila	$L_q = \frac{\lambda^2}{(\mu-\lambda)\mu}$
Tempo médio de espera de cliente no sistema	$W_s = \frac{1}{\mu-\lambda}$
Tempo médio de espera de cliente na fila	$W_q = \frac{\lambda}{(\mu-\lambda)\mu}$
Probabilidade de ocorrência do estado 0	$p_0 = 1 - \rho$
Probabilidade de ocorrência do estado n	$p_n = \rho^n(1-\rho)$
Probabilidade de existirem k ou mais clientes no sistema	$P(N \geq k) = \rho^k$
Probabilidade do tempo de espera na fila ser zero	$P(u_q = 0) = 1 - \rho$
Probabilidade do tempo de espera na fila ser maior que t	$P(u_q > t) = \frac{\lambda}{\mu} e^{-(\mu-\lambda)t}$
Probabilidade do tempo gasto no sistema ser maior que t	$P(u_s > t) = e^{-(\mu-\lambda)t}$

4.6.2 Modelo $(M/M/s) : (FIFO/\infty/\infty)$ para $(s > 1)$

Neste modelo, a única diferença com o modelo anterior é o número de servidores, pois estamos em presença de um modelo com vários servidores ($s > 1$). Então o modelo assume que o processo de chegada de clientes é um processo de Poisson, os tempos de serviço são independentes com uma distribuição exponencial, a disciplina de serviço é $FIFO$, a capacidade do sistema e a população são infinitos. Um exemplo deste modelo é o de um supermercado que tem vários postos de serviço ou ainda, pode-se consultar a aplicação deste modelo nos trabalhos [14], [50] e [51].

Este modelo pode ser representado através da Figura 4.2.

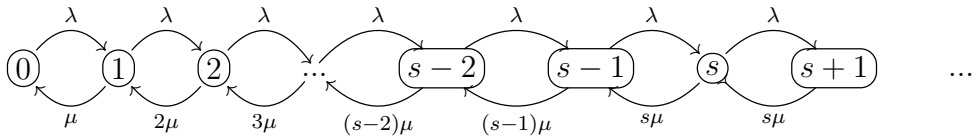


Figura 4.2: Modelo $(M/M/s) : (FIFO/\infty/\infty)$.

Temos assim um modelo em que a taxa de chegada de clientes é sempre igual a λ, ou seja,

$$\lambda_n = \lambda, \quad n = 0, 1, 2, ...$$

e a taxa de saída de clientes varia com o estado do sistema, ou seja,

$$\mu_n = \begin{cases} n\mu, & se \quad n = 1, 2, ..., s-1 \\ s\mu, & se \quad n = s, s+1, ... \end{cases} .$$

Como estamos em presença de múltiplos servidores ($s > 1$), então para estados onde os servidores não estejam todos ocupados a distribuição exponencial do intervalo de tempo entre saídas de clientes servidos tem parâmetros $k\mu$ (com $k < s$), onde k representa o número de servidores ocupados nesse estado.

Para estados onde os servidores estejam todos ocupados a distribuição exponencial do intervalo de tempo entre saídas de clientes servidos tem parâmetros $s\mu$.

Para determinarmos as distribuições estacionária devemos ter em conta o seguinte:

1) Estado onde os servidores não estejam todos ocupados ($0 \leq n \leq s-1$).

   Estado 0

   Usando a equaão (3.17) com $j = 0$ obtemos:

$$p_0 q_0 = \sum_{k \neq 0} p_k q_{k0} \Leftrightarrow p_0 q_0 = p_1 q_{10} \Leftrightarrow p_0 \lambda = p_1 \mu \Leftrightarrow p_1 = \frac{\lambda}{\mu} p_0, \tag{4.18}$$

   uma vez que $q_{k0} = 0$ se $k \neq 1$.

O raciocínio para achar as distribuições estacionária é igual ao modelo anterior, ou seja, a entrada no estado 0 ocorre quando o processo sai do estado 1 para o estado 0 devido a saída de um cliente do sistema (o fluxo de entradas no estado 0 é assim quantificado por μp_1) e a saída do estado 0 ocorre quando o processo sai do estado 0 para o estado 1 devido chegada de um cliente à fila (o fluxo de saídas do estado 0 é assim quantificado por λp_0).

Estado 1

Usando a equação (3.17) com $j = 1$, obtemos:

$$p_1 q_1 = \sum_{k \neq 1} p_k q_{k1} \quad \Leftrightarrow \quad p_1 q_1 = p_0 q_{01} + p_2 q_{21} \Leftrightarrow p_1(\lambda + \mu) = p_0\lambda + p_2 2\mu$$
$$\Leftrightarrow \quad p_2 = \frac{p_0\lambda - p_1(\lambda + \mu)}{2\mu},$$

substituindo o valor da equação (4.18) obtemos:

$$p_2 = \left(\frac{\lambda}{\mu}\right)^2 \frac{p_0}{2}. \tag{4.19}$$

uma vez que $q_{k1} = 0$, se $k \neq 0, 2$.

A entrada no estado 1 ocorre quando o processo sai do estado 0 para o estado 1 devido a chegada de um cliente a fila ou quando o processo sai do estado 2 para o estado 1 devido a saída de um cliente do sistema (o fluxo de entradas no estado 1 é assim quantificado por $\lambda p_0 + 2\mu p_2$) e a saída do estado 1 ocorre quando o processo sai do estado 1 para o estado 2 devido a chegada de um cliente à fila ou quando o processo sai do estado 1 para o esta 0 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $\lambda p_1 + \mu p_1$).

Estado 2

Usando a equação (3.17) com $j = 2$, obtemos:

$$p_2 q_2 = \sum_{k \neq 2} p_k q_{k2} \quad \Leftrightarrow \quad p_2 q_2 = p_1 q_{12} + p_3 q_{32} \Leftrightarrow p_2(\lambda + 2\mu) = p_1\lambda + p_3 3\mu$$
$$\Leftrightarrow \quad p_3 = \frac{p_2(\lambda + 2\mu) - p_1\lambda}{3\mu},$$

substituído os valores das equações (4.18) e (4.19) obtemos,

$$p_3 = \left(\frac{\lambda}{\mu}\right)^3 \frac{p_0}{6}. \tag{4.20}$$

Note-se que $q_{k2} = 0$ se $\forall k \neq 1, 3$.

A entrada no estado 2 ocorre quando o processo sai do estado 1 para o estado 2 devido a chegada de um cliente a fila ou quando o processo sai do estado 3 para o estado 2 devido a saída de um cliente do sistema (o fluxo de entradas no estado 2 é assim quantificado por $\lambda p_1 + 3\mu p_3$) e a saída do estado 2 ocorre quando o processo sai do estado 2 para o estado 3 devido a chegada de um cliente à fila ou quando o processo sai do estado 2 para o esta 1 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $\lambda p_2 + 2\mu p_2$).

Estado n

Pelo mesmo raciocínio dos estados anteriores, conclui-se que:

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} p_0 \text{ para } 0 \leq n \leq s-1.$$

2) Estado onde os servidores estejam todos ocupados ($n \geq s$). Supondo que $s = 3$, temos:

Estado 3

Usando a equação (3.17) com $j = 3$, obtemos:

$$p_3 q_3 = \sum_{k \neq 3} p_k q_{k3} \quad \Leftrightarrow \quad p_3 q_3 = p_2 q_{23} + p_4 q_{43} \Leftrightarrow p_3(\lambda + 3\mu) = p_2\lambda + p_4 4\mu$$
$$\Leftrightarrow \quad p_4 = \frac{p_3(\lambda + 3\mu) - p_2\lambda}{4\mu},$$

substituindo os valores das equações (4.19) e (4.20) obtemos:

$$p_4 = \left(\frac{\lambda}{\mu}\right)^4 \frac{p_0}{18}. \tag{4.21}$$

Note-se que $q_{k3} = 0$ se $\forall k \neq 2, 4$.

A entrada no estado 3 ocorre quando o processo sai do estado 2 para o estado 3 devido a chegada de um cliente a fila ou quando o processo sai do estado 4 para o estado 3 devido a saída de um cliente do sistema (o fluxo de entradas no estado 2 é assim quantificado por $\lambda p_2 + 4\mu p_4$) e a saída do estado 3 ocorre quando o processo sai do estado 3 para o estado 4 devido a chegada de um cliente à fila ou quando o processo sai do estado 3 para o esta 2 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $\lambda p_3 + 3\mu p_3$).

Estado 4

Usando a equação (3.17) com $j = 4$, obtemos:

$$p_4 q_4 = \sum_{k \neq 4} p_k q_{k4} \Leftrightarrow p_4 q_4 = p_3 q_{34} + p_5 q_{54} \Leftrightarrow p_4(\lambda + 4\mu) = p_3\lambda + p_5 5\mu$$
$$\Leftrightarrow p_5 = \frac{p_4(\lambda + 4\mu) - p_3\lambda}{5\mu},$$

substituindo os valores das equações (4.20) e (4.21) obtemos:

$$p_5 = \left(\frac{\lambda}{\mu}\right)^5 \frac{p_0}{54}. \tag{4.22}$$

Note-se que $q_{k4} = 0$ se $\forall k \neq 3, 5$.

A entrada no estado 4 ocorre quando o processo sai do estado 3 para o estado 4 devido a chegada de um cliente a fila ou quando o processo sai do estado 5 para o estado 4 devido a saída de um cliente do sistema (o fluxo de entradas no estado 4 é assim quantificado por $\lambda p_3 + 5\mu p_5$) e a saída do estado 4 ocorre quando o processo sai do estado 4 para o estado 5 devido a chegada de um cliente à fila ou quando o processo sai do estado 5 para o esta 4 por saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $\lambda p_4 + 5\mu p_5$).

Do mesmo modo, seguindo o raciocínio dos estados anteriores conclui-se que:

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{s! s^{n-s}} p_0 \text{ se } n \geq s.$$

Daí, p_n para o modelo em causa pode ser escrito da seguinte forma:

$$p_n = \begin{cases} \frac{(\frac{\lambda}{\mu})^n}{n!} p_0, & se \quad 0 \leq n \leq s-1 \\ \frac{(\frac{\lambda}{\mu})^n}{s! s^{n-s}} p_0, & se \quad n \geq s \end{cases},$$

consequentemente, se $\lambda < s\mu$ (de modo que a taxa de ocupação $\rho = \frac{\lambda}{s\mu} < 1$), então, p_0 é calculado a partir da seguinte equação:

$$\sum_{n=0}^{s-1} \frac{(\frac{\lambda}{\mu})^n}{n!} p_0 + \sum_{n=s}^{\infty} \frac{(\frac{\lambda}{\mu})^n}{s! s^{n-s}} p_0 = 1 \tag{4.23}$$

$$\begin{aligned} p_0 &= \left[\sum_{n=0}^{s-1}\frac{(\frac{\lambda}{\mu})^n}{n!}+\sum_{n=s}^{\infty}\frac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}\right]^{-1} \\ &= \left[\sum_{n=0}^{s-1}\frac{(\frac{\lambda}{\mu})^n}{n!}+\frac{(\frac{\lambda}{\mu})^s}{s!}\sum_{n=s}^{\infty}\left(\frac{\lambda}{s\mu}\right)^{n-s}\right]^{-1} \\ &= \left[\sum_{n=0}^{s-1}\frac{(\frac{\lambda}{\mu})^n}{n!}+\frac{(\frac{\lambda}{\mu})^s}{s!}\sum_{n=s}^{\infty}\rho^{n-s}\right]^{-1} \\ &= \left[\sum_{n=0}^{s-1}\frac{(\frac{\lambda}{\mu})^n}{n!}+\frac{(\frac{\lambda}{\mu})^s}{s!}\frac{1}{1-\rho}\right]^{-1}. \end{aligned}$$

Com estes resultados tendo em conta a equação (4.3), estamos em condições de determinar o número médio de clientes na fila, isto é,

$$L_q=\sum_{n=s+1}^{\infty}(n-s)p_n=\sum_{j=1}^{\infty}jp_{j+s}=\sum_{j=1}^{\infty}j\frac{(\frac{\lambda}{\mu})^s}{s!}\rho^j p_0=p_0\frac{(\frac{\lambda}{\mu})^s}{s!}\rho\frac{d}{d\rho}\left(\sum_{j=1}^{\infty}\rho^j\right)=\frac{p_0(\frac{\lambda}{\mu})^s\rho}{s!(1-\rho)^2}; \tag{4.24}$$

Usando a equação (4.24), juntamente com as equações (4.6), (4.8) e (4.7), podemos determinar:

- o valor de W_q,
$$W_q=\frac{L_q}{\lambda};$$
- o valor de L_s,
$$L_s=L_q+\frac{\lambda}{s\mu};$$
- o valor de W_s,
$$W_s=W_q+\frac{1}{\mu}.$$

Para achar a distribuição de probabilidade do tempo na fila no caso de modelos com múltiplos servidores, pode-se generalizar o caso visto em modelos de único servidor. Neste caso devemos ter em conta o seguinte:

- No caso em que os servidores não estejam todos ocupas $0 \leq n \leq s-1$, o cliente que acaba de chegar ao sistema é atendido imediatamente, ou seja:
$$P(u_q=0)=\sum_{n=0}^{s-1}p_n.$$
- No caso em que os servidores estejam todos ocupados $n \geq s$, os intervalos entre as saídas sucessivas são independentes e identicamente distribuídos com

distribuição exponencial com parâmetro $s\mu$ e o tempo total até a saída $n-s+1$ tem distribuição gama de parâmetro $(s\mu, n-s+1)$, ou seja,

$$\begin{aligned}
P(0 < u_q \leq t) &= \sum_{n=s}^{\infty} P(0 < u_q \leq t | N = n) P(N = n) \\
&= \sum_{n=s}^{\infty} \left[\int_0^t \frac{(s\mu t)^{n-s}}{(n-s)!} s\mu e^{-s\mu x} dx \right] p_n \\
&= \int_0^t s\mu e^{-s\mu x} \left[\sum_{n=s}^{\infty} \frac{(s\mu t)^{n-s}}{(n-s)!} \frac{\left(\frac{\lambda}{\mu}\right)^n}{s! s^{n-s}} p_0 \right] dx \\
&= \int_0^t s\mu e^{-s\mu x} \left[\frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} p_0 e^{\lambda x} \right] dx = \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} p_0 s\mu \int_0^t e^{-s\mu x + \lambda x} dx \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^s}{(s\mu - \lambda) s!} p_0 s\mu \left[1 - e^{-(s\mu - \lambda)t} \right].
\end{aligned}$$

Portanto,

$$\begin{aligned}
P(u_q \leq t) &= P(u_q = 0) + P(0 < u_q \leq t),\ t \geq 0 \\
&= \sum_{n=0}^{s-1} p_n + \frac{\left(\frac{\lambda}{\mu}\right)^s}{(s\mu - \lambda) s!} p_0 s\mu \left[1 - e^{-(s\mu - \lambda)t} \right] \\
&= 1 - \frac{\left(\frac{\lambda}{\mu}\right)^s p_0}{(s\mu - \lambda) s!} e^{-(s\mu - \lambda)t},
\end{aligned}$$

entao, a probabilidade do tempo na fila ser maior t será:

$$\begin{aligned}
P(u_q > t) &= 1 - P(u_q \leq t) \\
&= 1 - 1 + \frac{\left(\frac{\lambda}{\mu}\right)^s p_0}{(s\mu - \lambda) s!} e^{-(s\mu - \lambda)t} \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^s p_0}{(s\mu - \lambda) s!} e^{-(s\mu - \lambda)t},
\end{aligned}$$

usando a equação (4.16) obtemos:

$$P(u_s > t) = \frac{P(u_q > t)}{\rho}.$$

Tabela 4.2: Tabela das características do modelo $(M/M/s) : (FIFO/\infty/\infty)$

Chegada : Poissoneana Taxa : λ clientes/unidade de tempo População = ∞ Nº máximo de clientes na fila = ∞	Tempo de atendimento: exponencial Taxa : μ clientes/un. tempo e posto de serviço Nº de servidores = s Taxa de ocupação $\rho = \frac{\lambda}{s\mu}$ com $\rho < 1$ Taxa de desocupação = 1 - ρ
Número médio de clientes no sistema	$L_s = L_q + \frac{\lambda}{s\mu}$
Número médio de clientes que aguardam na fila	$L_q = \frac{p_0(\frac{\lambda}{\mu})^s\rho}{s!(1-\rho)^2}$
Tempo médio de espera de cliente no sistema	$W_s = W_q + \frac{1}{\mu}$
Tempo médio de espera de cliente na fila	$W_q = \frac{L_q}{\lambda}$
Probabilidade de ocorrência do estado 0	$p_0 = \left[\sum_{n=0}^{s-1} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!}\cdot\frac{1}{1-\rho}\right]^{-1}$
Probabilidade de ocorrência do estado n	$p_n = \begin{cases} \frac{(\frac{\lambda}{\mu})^n}{n!}p_0, & se \quad 0 \leq n \leq s-1 \\ \frac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}p_0, & se \quad n \geq s \end{cases}$
Probabilidade do tempo de espera na fila ser zero	$P(u_q = 0) = \sum_{n=0}^{s-1} p_n$
Probabilidade do tempo de espera na fila ser maior que t	$P(u_q > t) = \frac{\left(\frac{\lambda}{\mu}\right)^s p_0}{(s\mu-\lambda)s!}e^{-(s\mu-\lambda)t}$
Probabilidade do tempo gasto no sistema ser maior que t	$P(u_s > t) = \frac{P(u_q>t)}{\rho}$

4.6.3 Modelo $(M/M/1) : (FIFO/k/\infty)$

Este modelo assume que a distribuição dos tempos entre chegadas e os tempos de serviço é exponencial, tem único servidor, o número máximo de clientes no sistema é k, a disciplina de serviço é $FIFO$ (primeiro cliente a chegar primeiro a ser servido e a sair) e a população é infinita. Um exemplo deste modelo é de um posto de identificação no qual tem um número limitado para k clientes e um único posto de serviço.

Este modelo pode ser representado através da Figura 4.3.

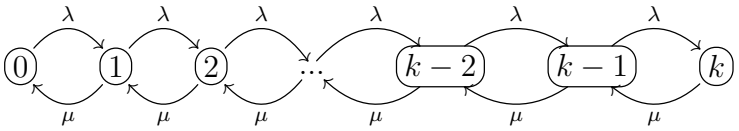


Figura 4.3: Modelo $(M/M/1) : (FIFO/k/\infty)$.

Sabendo que o número máximo de cliente no sistema é k, isto significa que qualquer cliente que chega depois do sistema estar completo é recusado a entrar. Assim, a probabilidade do sistema estar num estado $n \geq k+1$ é zero.

Temos assim um modelo em que a taxa de chegada de clientes λ_n depende do

estado do sistema

$$\lambda_n = \begin{cases} \lambda, & se \quad n = 0, 1, 2, ..., k-1 \\ 0, & se \quad n \geq k \end{cases} .$$

Neste caso, como a taxa de chegada de clientes depende do estado do sistema, então devemos usar $\overline{\lambda}$ (média ponderada das taxas λ_n), onde

$$\overline{\lambda} = \sum_{n=0}^{k-1} \lambda p_n = \lambda \sum_{n=0}^{k-1} p_n = \lambda(1 - p_k) \tag{4.25}$$

e a taxa de saída de clientes será:

$$\mu_n = \begin{cases} \mu, & se \quad n = 0, 1, 2, ..., k \\ 0, & se \quad n > k, \end{cases}$$

e a taxa de ocupação será igual a $\frac{\overline{\lambda}}{\mu}$.

De seguida determinamos as distribuições para cada estado. Lembrar ainda, que os cálculos deste modelo tem o mesmo raciocínio do modelo $(M/M/1) : (FIFO/\infty/\infty)$

Estado 0

Usando a equação (3.17) com $j = 0$, obtemos::

$$\lambda p_0 = \mu p_1 \Leftrightarrow p_1 = \frac{\lambda}{\mu}.$$

uma vez que $q_{k0} = 0$ se $k \neq 1$.

Estado 1

Usando a equação (3.17) com $j = 1$, obtemos:

$$\begin{aligned} \lambda p_0 + \mu p_2 &= \lambda p_1 + \mu p_1 \\ p_2 &= \frac{p_1(\lambda + \mu) - \lambda p_0}{\mu} = \left(\frac{\lambda}{\mu}\right)^2 p_0. \end{aligned}$$

uma vez que $q_{k1} = 0$ se $k \neq 0, 2$.

Estado 2

Usando a equação (3.17) com $j = 2$, obtemos:

$$\begin{aligned} \lambda p_1 + \mu p_3 &= \mu p_2 + \lambda p_2 \\ p_3 &= \frac{\mu p_2 + \lambda p_2 - \lambda p_1}{\mu} = \left(\frac{\lambda}{\mu}\right)^3 p_0. \end{aligned}$$

uma vez que $q_{k2} = 0$ se $k \neq 1, 3$.

Estado n

Para o estado n segue-se o mesmo raciocínio, concluindo-se que

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \Leftrightarrow p_n = \rho^n p_0,$$

onde $\rho = \frac{\lambda}{\mu}$, ou seja, p_n também pode ser representado da seguinte maneira:

$$p_n = \begin{cases} \rho^n p_0, & se \quad n = 0, 1, 2, ..., k \\ 0, & se \quad n > k \end{cases},$$

e p_0 é obtido a partir da seguinte equação:

$$\begin{aligned} \sum_{n=0}^{k} \rho^n p_0 &= 1 \ (\rho \neq 1) \\ p_0 &= \left[\sum_{n=0}^{k} \rho^n\right]^{-1} \\ &= \left[\frac{1-\rho^{k+1}}{1-\rho}\right]^{-1} \\ &= \frac{1-\rho}{1-\rho^{k+1}}, \end{aligned}$$

usando as equações (4.2), (4.8), (4.5) e (4.6), podemos determinar:

- o valor de L_s,

$$\begin{aligned} L_s &= \sum_{n=0}^{k} n p_n = \frac{1-\rho}{1-\rho^{k+1}} \rho \sum_{n=1}^{k} \frac{d}{d\rho}\left(\rho^n\right) \\ &= \frac{1-\rho}{1-\rho^{k+1}} \rho \frac{d}{d\rho}\left(\sum_{n=1}^{k} \rho^n\right) = \frac{1-\rho}{1-\rho^{k+1}} \rho \frac{d}{d\rho}\left(\frac{1-\rho^{k+1}}{1-\rho}\right) \\ &= \frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}; \end{aligned}$$

- o valor de L_q,

$$L_q = L_s - \frac{\overline{\lambda}}{\mu};$$

- o valor de W_s,

$$W_s = \frac{L_s}{\overline{\lambda}};$$

- o valor de W_q,

$$W_q = \frac{L_q}{\overline{\lambda}},$$

e

$$P(u_q = 0) = p_0. \tag{4.26}$$

Tabela 4.3: Tabela das características do modelo $(M/M/1) : (FIFO/k/\infty)$

Chegada : Poissoneana Taxa : $\lambda_n = \begin{cases} \lambda, & se \quad n = 0,1,2,...,k \\ 0, & se \quad n > k \end{cases}$	Tempo de atendimento: exponencial Taxa: $\mu_n = \begin{cases} \mu, & se \quad n = 0,1,2,...,k \\ 0, & se \quad n > k \end{cases}$
Média ponderada das taxas : $\overline{\lambda} = \lambda(1 - p_k)$ População $= \infty$ Nº máximo de clientes no sistema $= k$ Nº máximo de clientes na fila $= k - s$	Nº de servidores $= 1$, $\rho = \frac{\lambda}{\mu}$ Taxa de ocupação: $\frac{\overline{\lambda}}{\mu}$ com $\frac{\overline{\lambda}}{\mu} < 1$ Taxa de desocupação$=1 - \frac{\overline{\lambda}}{\mu}$
Número médio de clientes no sistema Número médio de clientes que aguardam na fila Tempo médio de espera de cliente no sistema Tempo médio de espera de cliente na fila Probabilidade de ocorrência do estado 0 Probabilidade de ocorrência do estado n	$L_s = \frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}$ $L_q = L_s - \frac{\overline{\lambda}}{\mu}$ $W_s = \frac{L_s}{\overline{\lambda}}$ $W_q = \frac{L_q}{\overline{\lambda}}$ $p_0 = \frac{1-\rho}{1-\rho^{k+1}}$ $p_n = \begin{cases} \rho^n p_0, & se \quad n = 0,1,2,...,k \\ 0, & se \quad n > k \end{cases}$
Probabilidade do tempo de espera na fila ser zero	$P(u_q = 0) = p_0$

4.6.4 Modelo $(M/M/s) : (FIFO/k/\infty)$ para $(1 < s \leq k)$

Este modelo é semelhante ao modelo $(M/M/s) : (FIFO/\infty/\infty)$, a única diferença é que este modelo refere-se a um número k (finito) de clientes no sistema, ou seja, o $(k+1)$ - ésimo cliente que acaba de chegar já não entra no sistema. Um exemplo deste modelo pode ser um posto de atendimento do SEF que contém múltiplos servidores e o sistema tem uma limitação de k clientes. Para mais detalhes sobre a aplicação deste modelo pode-se consultar [43], [42].

Este modelo pode ser representado através da Figura 4.4.

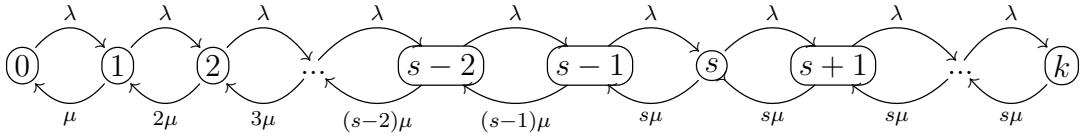


Figura 4.4: Modelo $(M/M/s) : (FIFO/k/\infty)$.

Temos assim um modelo em que a taxa de chegada de clientes λ_n e a taxa de saída de clientes μ_n, dependem do estado do sistema, ou seja,

$$\lambda_n = \begin{cases} \lambda, & se \quad n = 0,1,2,...,k-1 \\ 0, & se \quad n \geq k \end{cases}$$

e

$$\mu_n = \begin{cases} n\mu, & se \quad n = 0, 1, 2, ..., s-1 \\ s\mu, & se \quad n = s, ..., k \\ 0, & se \quad n > k, \end{cases},$$

então, para os cálculos das distribuições de cada estado o raciocínio é igual ao do modelo $(M/M/s) : (FIFO/\infty/\infty)$, ou seja:

$$p_n = \begin{cases} \frac{(\frac{\lambda}{\mu})^n}{n!} p_0, & se \quad n = 0, 1, 2, ..., s-1 \\ \frac{(\frac{\lambda}{\mu})^n}{s! s^{n-s}} p_0, & se \quad n = s, s+1, ..., k \\ 0, & se \quad n > k, \end{cases},$$

onde, p_0 é calculado a partir da seguinte equação:

$$\sum_{n=0}^{s-1} \frac{(\frac{\lambda}{\mu})^n}{n!} p_0 + \sum_{n=s}^{k} \frac{(\frac{\lambda}{\mu})^n}{s! s^{n-s}} p_0 = 1 \Leftrightarrow p_0 = \left[\sum_{n=0}^{s-1} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!} \sum_{n=s}^{k} \left(\frac{\lambda}{s\mu} \right)^{n-s} \right]^{-1} \quad (4.27)$$

Como $\rho = \frac{\lambda}{s\mu}$, então o segundo somatório da equação (4.27) fica:

$$\sum_{n=s}^{k} \rho^{n-s} = \begin{cases} \frac{1-\rho^{k-s+1}}{1-\rho}, & se \quad \rho \neq 1 \\ k-s+1, & se \quad \rho = 1 \end{cases},$$

portanto,

$$p_0 = \begin{cases} \left[\sum_{n=0}^{s-1} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!} \left(\frac{1-\rho^{k-s+1}}{1-\rho} \right) \right]^{-1}, & se \quad \rho \neq 1 \\ \left[\sum_{n=0}^{s-1} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!} (k-s+1) \right]^{-1}, & se \quad \rho = 1 \end{cases},$$

Utilizando a equação (4.3) obtemos o número médio de clientes na fila para $\rho \neq 1$.

$$\begin{aligned} L_q &= \sum_{n=s+1}^{k} (n-s) p_n = \sum_{n=s+1}^{k} (n-s) \frac{(\frac{\lambda}{\mu})^n}{s! s^{n-s}} p_0 \\ &= \sum_{n=s+1}^{k} (n-s) \frac{\rho^n s^s}{s!} p_0 = \frac{\rho^{s+1} s^s}{s!} p_0 \sum_{n=s+1}^{k} (n-s) \rho^{n-s-1} \\ &= \frac{\rho^{s+1} s^s}{s!} p_0 \sum_{i=1}^{k-s} i \rho^{i-1} = \frac{\rho^{s+1} s^s}{s!} p_0 \frac{d}{d\rho} \left(\sum_{i=0}^{k-s} \rho^i \right) \\ &= \frac{\rho^{s+1} s^s}{s!} p_0 \frac{d}{d\rho} \left(\frac{1-\rho^{k-s+1}}{1-\rho} \right) \end{aligned}$$

$$= \frac{P_0 (\frac{\lambda}{\mu})^s \rho}{s!(1-\rho)^2} \left[1 - \rho^{k-s} - (k-s)\rho^{k-s}(1-\rho) \right]. \quad (4.28)$$

Para $\rho = 1$ obtemos:

$$\begin{aligned} L_q &= \sum_{n=s+1}^{k} (n-s)p_n = \sum_{n=s+1}^{k} (n-s)\frac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}p_0 \\ &= \sum_{n=s+1}^{k} (n-s)\frac{s^s\rho^n p_0}{s!} = \frac{s^s p_0}{s!}\sum_{n=s+1}^{k}(n-s) \end{aligned}$$

$$= \frac{s^s p_0}{s!}\sum_{i=1}^{k-s} i = \frac{s^s p_0}{s!}\frac{(k-s)(k-s+1)}{2} \tag{4.29}$$

O valor médio de clientes no sistema para $\rho \neq 1$ é

$$\begin{aligned} L_s &= L_q + \frac{\overline{\lambda}}{\mu} = L_q + \frac{\lambda\left(1 - \frac{(\frac{\lambda}{\mu})^k}{s!s^{k-s}}p_0\right)}{\mu} \\ &= \frac{s^s\rho^{s+1}p_0}{s!(1-\rho)^2}\left[1 - \rho^{k-s}(1+(1-\rho)(k-s\rho))\right] + s\rho; \end{aligned}$$

para $\rho = 1$ é

$$L_s = L_q + \frac{\overline{\lambda}}{\mu} = L_q + \frac{\lambda\left(1 - \frac{(\frac{\lambda}{\mu})^k}{s!s^{k-s}}p_0\right)}{\mu} = \frac{s^s p_0}{s!}\left[\frac{(k-s)(k-s+1)-s}{2}\right] + s. \tag{4.30}$$

Os tempos médios que um cliente aguarda na fila ou no sistema são obtidos a partir das equações

$$W_q = \frac{L_q}{\overline{\lambda}} \qquad \text{e} \qquad W_s = \frac{Ls}{\overline{\lambda}}, \tag{4.31}$$

ou seja,

$$W_q = \begin{cases} \frac{s^{s-1}\rho^s p_0}{\mu()^2(s!-s^s\rho^k p_0)}\left[1-\rho^{k-s}(1+(1-\rho)(k-s))\right], & se \quad \rho \neq 1 \\ \frac{s^{s-1}p_0(k-s)(k-s+1)}{2\mu(s!-sp_0)}, & se \quad \rho = 1 \end{cases},$$

e

$$W_s = \begin{cases} \frac{s^{s-1}\rho^s p_0}{\mu()^2(s!-s^s\rho^k p_0)}\left[1-\rho^{k-s}(1+(1-\rho)(k-s))\right] + \frac{1}{\mu}, & se \quad \rho \neq 1 \\ \frac{s^{s-1}p_0(k-s)(k-s+1)}{2\mu(s!-sp_0)} + \frac{1}{\mu}, & se \quad \rho = 1 \end{cases}.$$

Onde $\overline{\lambda}$ é obtido através da equação (4.25).

A probabilidade do tempo de espera na fila ser zero será;

$$P(u_q = 0) = \sum_{n=0}^{s-1} p_n.$$

Tabela 4.4: Características do modelo $(M/M/s) : (FIFO/k/\infty)$

Chegada : Poissoneana Taxa :	Tempo de atendimento: exponencial Taxa :
$\lambda_n = \begin{cases} \lambda, & se \quad n = 0,1,2,...,k-1 \\ 0, & se \quad n \geq k \end{cases}$	$\mu_n = \begin{cases} n\mu, & se \quad n = 0,1,2,...,s-1 \\ s\mu, & se \quad n = s,...,k \\ 0, & se \quad n > k \end{cases}$
Média ponderada das taxas : $\overline{\lambda} = \lambda(1-p_k)$	
População $= \infty$	Nº de servidores $= s$ $(s > 1)$, $\rho = \frac{\lambda}{s\mu}$
Nº máximo de clientes no sistema $= k$	Taxa de ocupação $= \frac{\overline{\lambda}}{s\mu}$
Nº máximo de clientes na fila $= k - s$	Taxa de desocupação $= 1 - \frac{\overline{\lambda}}{s\mu}$
Número médio de clientes no sistema	$L_s = \frac{s^s\rho^{s+1}p_0}{s!(1-\rho)^2}\left[1-\rho^{k-s}(1+(1-\rho)(k-s\rho))\right] + s\rho,\ \rho \neq 1$
Número médio de clientes que aguardam na fila	$L_q = \frac{P_0(\frac{\lambda}{\mu})^s\rho}{s!(1-\rho)^2}\left[1-\rho^{k-s}-(k-s)\rho^{k-s}(1-\rho)\right],\ \rho \neq 1$
Tempo médio de espera de cliente no sistema	$W_s = \frac{L_s}{\overline{\lambda}}$
Tempo médio de espera de cliente na fila	$W_q = \frac{L_q}{\overline{\lambda}}$
Probabilidade de ocorrência do estado 0	$p_0 = \left[\sum_{n=0}^{s-1}\frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!}\left(\frac{1-\rho^{k-s+1}}{1-\rho}\right)\right]^{-1},\ \rho \neq 1$
Probabilidade de ocorrência do estado n	$p_n = \begin{cases} \frac{(\frac{\lambda}{\mu})^n}{n!}p_0, & n = 0,1,2,...,s-1 \\ \frac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}p_0, & n = s,s+1,...,k \\ 0, & se \quad n > k \end{cases}$
Probabilidade do tempo de espera na fila ser zero	$P(u_q = 0) = \sum_{n=0}^{s-1} p_n$

4.6.5 Modelo $(M/M/1) : (FIFO/\infty/N)$

Este modelo assume que os tempos entre chegadas são independentes e identicamente distribuídos de acordo com uma distribuição exponencial, ou seja, o processo de entrada é Poisson, os tempos de serviço são independente e identicamente distribuídos de acordo com outra distribuição exponencial, existe um único servidor, a disciplina de serviço é $FIFO$ (primeiro a chegar primeiro a ser servido e a sair), a capacidade do sistema é infinita e tamanho da população é N (finito). Um exemplo da aplicação deste modelo pode-se consultar em [28].

Este modelo pode ser representado através da Figura 4.5.

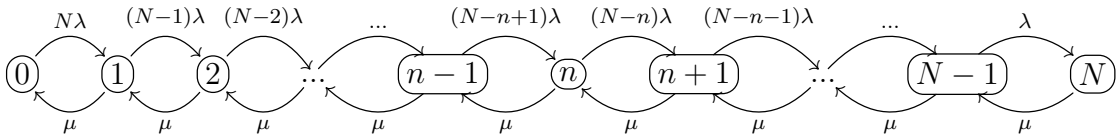


Figura 4.5: Modelo $(M/M/1) : (FIFO/\infty/N)$.

Se o número de clientes no sistema for igual a n $(n = 0,1,2,...,N)$, restam apenas

$N - n$ clientes na população.

Temos assim um modelo em que a taxa de chegada de clientes é dada por

$$\lambda_n = \begin{cases} (N-n)\lambda, & se \quad n = 0, 1, 2, ..., N \\ 0, & se \quad n > N \end{cases},$$

e a taxa de saída de clientes é dada por

$$\mu_n = \begin{cases} \mu, & se \quad n = 1, 2, ..., N \\ 0, & se \quad n > N \end{cases}.$$

Daí, determina-se as distribuições estacionária para cada estado, usando o mesmo raciocínio dos modelos anteriores.

Estado 0

Usando a equação (3.17) com $j = 0$ obtemos:

$$p_0 q_0 = \sum_{k \neq 0} p_k q_{k0} \Leftrightarrow p_0 q_0 = p_1 q_{10} \Leftrightarrow p_0 N\lambda = p_1 \mu \Leftrightarrow p_1 = \frac{N\lambda}{\mu} p_0, \tag{4.32}$$

uma vez que $q_{k0} = 0$ se $k \neq 1$.

A entrada no estado 0 ocorre quando o processo sai do estado 1 para o estado 0 devido a saída de um cliente no sistema (o fluxo de entradas no estado 0 é então quantificado por μp_1) e a saída do estado 0 ocorre quando o processo sai do estado 0 para o estado 1 devido a chegada de um cliente à fila (o fluxo de entradas no estado 0 é então quantificado por $N\lambda p_0$).

Estado 1

Usando a equação (3.17) com $j = 1$, obtemos:

$$\begin{aligned} p_1 q_1 = \sum_{k \neq 1} p_k q_{k1} \quad &\Leftrightarrow \quad p_1 q_1 = p_0 q_{01} + p_2 q_{21} \Leftrightarrow p_1((N-1)\lambda + \mu) = p_0 N\lambda + p_2 \mu \\ &\Leftrightarrow \quad p_2 = \frac{p_1((N-1)\lambda + \mu) - p_0 N\lambda}{\mu}, \end{aligned}$$

substituindo o valor da equação (4.32) obtemos:

$$p_2 = N(N-1)\left(\frac{\lambda}{\mu}\right)^2 p_0. \tag{4.33}$$

Note-se que $q_{k1} = 0$ se $\forall k \neq 0, 2$.

A entrada no estado 1 ocorre quando o processo sai do estado 0 para o estado 1 devido a chegada de um cliente a fila ou quando o processo sai do estado 2 para o estado 1 devido a saída de um cliente do sistema (o fluxo de entradas no estado 1 é assim quantificado por $N\lambda p_0 + \mu p_2$) e a saída do estado 1 ocorre quando o processo sai do estado 1 para o estado 2 devido a chegada de um cliente à fila ou quando o processo sai do estado 1 para o esta 0 devido a saída de um cliente

do sistema (o fluxo de saídas do estado 1 é assim quantificado por $(N-1)\lambda p_1+\mu p_1$).

Estado 2

Usando a equação (3.17) com $j=2$, obtemos:

$$p_2q_2=\sum_{k\neq 2}p_kq_{k2} \Leftrightarrow p_2q_2=p_1q_{12}+p_3q_{32} \Leftrightarrow p_2((N-2)\lambda+\mu)=p_1(N-1)\lambda+p_3\mu$$
$$\Leftrightarrow p_3=\frac{p_2((N-2)\lambda+\mu)-p_1(N-1)\lambda}{\mu},$$

substituído os valores das equações (4.32) e (4.33) obtemos:

$$p_3=N(N-1)(N-2)\left(\frac{\lambda}{\mu}\right)^3 p_0. \tag{4.34}$$

Note-se que $q_{k2}=0$ se $\forall k\neq 1,3$.

A entrada no estado 2 ocorre quando o processo sai do estado 1 para o estado 2 devido a chegada de um cliente a fila ou quando o processo sai do estado 3 para o estado 2 devido a saída de um cliente do sistema (o fluxo de entradas no estado 2 é assim quantificado por $(N-1)\lambda p_1+\mu p_3$) e a saída do estado 2 ocorre quando o processo sai do estado 2 para o estado 3 devido a chegada de um cliente à fila ou quando o processo sai do estado 2 para o esta 1 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $(N-2)\lambda p_2+\mu p_2$).

Estado 3

Usando a equação (3.17) com $j=3$, obtemos:

$$p_3q_3=\sum_{k\neq 3}p_kq_{k3} \Leftrightarrow p_3q_3=p_2q_{23}+p_4q_{43} \Leftrightarrow p_3((N-3)\lambda+\mu)=p_2(N-2)\lambda+p_4\mu$$
$$\Leftrightarrow p_4=\frac{p_3((N-3)\lambda+\mu)-p_2(N-2)\lambda}{\mu},$$

substituindo os valores das equações (4.33) e (4.34) obtemos:

$$p_4=N(N-1)(N-2)(N-n+3)\left(\frac{\lambda}{\mu}\right)^4 p_0. \tag{4.35}$$

Note-se que $q_{k3}=0$ se $\forall k\neq 2,4$.

A entrada no estado 3 ocorre quando o processo sai do estado 2 para o estado 3 devido a chegada de um cliente a fila ou quando o processo sai do estado 4 para o estado 3 devido a saída de um cliente do sistema (o fluxo de entradas no estado 2 é assim quantificado por $(N-2)\lambda p_2+\mu p_4$) e a saída do estado 3 ocorre quando o processo sai do estado 3 para o estado 4 devido a chegada de um cliente à fila ou quando o processo sai do estado 3 para o esta 2 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $(N-3)\lambda p_3+\mu p_3$).

Estado n

Para o estado n, concluímos também que

$$p_n = \begin{cases} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n p_0, & se \quad n = 1, 2, ..., N \\ 0, & se \quad n > N \end{cases},$$

onde p_0 é obtido a partir da seguinte equação:

$$\begin{aligned} \sum_{n=0}^{N} p_n &= 1 \Leftrightarrow \sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n p_0 = 1 \\ p_0 &= \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}. \end{aligned}$$

Usando as equações (4.2), (4.3), (4.5) e (4.6) obtemos:

- o valor de L_s,
$$L_s = \sum_{n=0}^{N} n p_n.$$
- o valor de L_q,
$$L_q = \sum_{n=2}^{N} (n-1) p_n.$$
- o valor de W_s,
$$W_s = \frac{L_s}{\overline{\lambda}}.$$
- o valor de W_q,
$$W_q = \frac{L_q}{\overline{\lambda}}.$$

onde,

$$\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n p_n = \sum_{n=0}^{N} (N-n)\lambda p_n = \lambda(N - L_s). \tag{4.36}$$

4.6.6 Modelo $(M/M/s) : (FIFO/\infty/N)$ para $(s > 1)$

Este modelo assume que os tempos entre chegadas são independentes identicamente distribuídos de acordo com uma distribuição exponencial, ou seja, o processo de entrada é Poisson, os tempos de serviço são independente identicamente distribuídos de acordo com outra distribuição exponencial, o número de servidor é s $(s > 1)$, a disciplina de serviço é $FIFO$ (primeiro a chegar primeiro a ser servido e a sair), a capacidade do sistema é infinita e tamanho da população é N (finito). Um exemplo da aplicação deste modelo pode-se consultar em [27].

Tabela 4.5: Características do modelo $(M/M/1) : (FIFO/\infty/N)$

Chegada : Poissoneana Taxa : $\lambda_n = \begin{cases} (N-n)\lambda, & se \quad n = 0,1,2,...,N \\ 0, & se \quad n \geq N \end{cases}$	Tempo de atendimento: exponencial Taxa : $\mu_n = \mu, para n = 1,2,...$
Média ponderada das taxa λ : $\overline{\lambda} = \lambda(N - L_s)$ População $= N$	Nº de servidores $= 1$ Taxa de ocupação $= \frac{\overline{\lambda}}{\mu}$ com $\frac{\overline{\lambda}}{\mu} < 1$ Taxa de desocupação $= 1 - \frac{\overline{\lambda}}{\mu}$
Número médio de clientes no sistema	$L_s = N - \frac{\mu}{\lambda}(1 - p_o)$
Número médio de clientes que aguardam na fila	$L_q = N - \frac{\lambda+\mu}{\lambda}(1 - p_0)$
Tempo médio de espera de cliente no sistema	$W_s = \frac{L_s}{\overline{\lambda}}$
Tempo médio de espera de cliente na fila	$W_q = \frac{L_q}{\overline{\lambda}}$
Probabilidade de ocorrência do estado 0	$p_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$
Probabilidade de ocorrência do estado n	$p_n = \begin{cases} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n p_0, & se \quad n = 1,2,...,N \\ 0, & se \quad n > N \end{cases}$

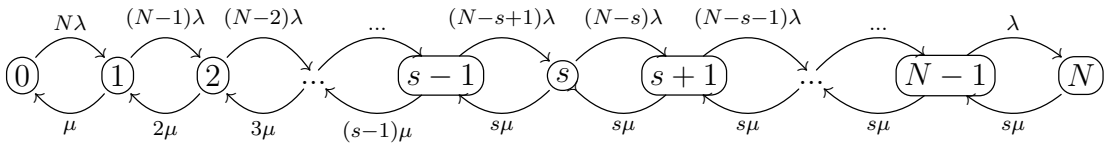


Figura 4.6: Modelo $(M/M/s) : (FIFO/\infty/N)$.

Este modelo pode ser representado através da Figura 4.6.

Temos assim um modelo em que a taxa de chegada de clientes é dada por,

$$\lambda_n = \begin{cases} (N-n)\lambda, & se \quad n = 0,1,2,...,N \\ 0, & se \quad n > N \end{cases},$$

a taxa de saída de clientes é dada por,

$$\mu_n = \begin{cases} n\mu, & se \quad n = 1,2,...,s-1 \\ s\mu, & se \quad n = s, s+1, ..., N \\ 0, & se \quad n > N \end{cases},$$

neste caso, as distribuições estacionária de é determinada pelo mesmo raciocínio referido nos modelos anteriores.

Estado 0

Usando a equação (3.17) com $j = 0$ obtemos:

$$p_0 q_0 = \sum_{k \neq 0} p_k q_{k0} \Leftrightarrow p_0 q_0 = p_1 q_{10} \Leftrightarrow p_0 N\lambda = p_1 \mu \Leftrightarrow p_1 = \frac{N\lambda}{\mu} p_0, \tag{4.37}$$

uma vez que $q_{k0} = 0$ se $k \neq 1$.

O raciocínio para discrição das distribuições estacionária é igual ao modelo anterior.

Estado 1

Usando a equação (3.17) com $j = 1$, obtemos:

$$\begin{aligned} p_1 q_1 = \sum_{k \neq 1} p_k q_{k1} \quad &\Leftrightarrow \quad p_1 q_1 = p_0 q_{01} + p_2 q_{21} \Leftrightarrow p_1((N-1)\lambda + \mu) = p_0 N\lambda + p_2 2\mu \\ &\Leftrightarrow \quad p_2 = \frac{p_1((N-1)\lambda + \mu) - p_0 N\lambda}{2\mu}, \end{aligned}$$

substituindo o valor da equação (4.37) obtemos:

$$p_2 = \frac{N(N-1)}{2} \left(\frac{\lambda}{\mu}\right)^2 p_0. \tag{4.38}$$

Note-se que $q_{k1} = 0$ se $\forall k \neq 0, 2$.

A entrada no estado 1 ocorre quando o processo sai do estado 0 para o estado 1 devido a chegada de um cliente a fila ou quando o processo sai do estado 2 para o estado 1 devido a saída de um cliente do sistema (o fluxo de entradas no estado 1 é assim quantificado por $N\lambda p_0 + 2\mu p_2$) e a saída do estado 1 ocorre quando o processo sai do estado 1 para o estado 2 devido a chegada de um cliente à fila ou quando o processo sai do estado 1 para o esta 0 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $(N-1)\lambda p_1 + \mu p_1$).

Estado 2

Usando a equação (3.17) com $j = 2$, obtemos:

$$\begin{aligned} p_2 q_2 = \sum_{k \neq 2} p_k q_{k2} \quad &\Leftrightarrow \quad p_2 q_2 = p_1 q_{12} + p_3 q_{32} \Leftrightarrow p_2((N-2)\lambda + 2\mu) = p_1(N-1)\lambda + p_3 3\mu \\ &\Leftrightarrow \quad p_3 = \frac{p_2((N-2)\lambda + 2\mu) - p_1(N-1)\lambda}{3\mu}, \end{aligned}$$

substituindo os valores das equações (4.37) e (4.38) obtemos:

$$p_3 = \frac{N(N-1)(N-2)}{6} \left(\frac{\lambda}{\mu}\right)^3 p_0. \tag{4.39}$$

Note-se que $q_{k2} = 0$ se $\forall k \neq 1, 3$.

A entrada no estado 2 ocorre quando o processo sai do estado 1 para o estado 2 devido a chegada de um cliente a fila ou quando o processo sai do estado 3 para o estado 2 devido a saída de um cliente do sistema (o fluxo de entradas no estado 2 é assim quantificado por $(N-1)\lambda p_1 + 3\mu p_3$) e a saída do estado 2 ocorre quando o processo sai do estado 2 para o estado 3 devido a chegada de um cliente à fila ou quando o processo sai do estado 2 para o esta 1 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $(N-2)\lambda p_2 + 2\mu p_2$).

Estado 3

Usando a equação (3.17) com $j = 3$, obtemos:

$$p_3 q_3 = \sum_{k \neq 3} p_k q_{k3} \quad \Leftrightarrow \quad p_3 q_3 = p_2 q_{23} + p_4 q_{43} \Leftrightarrow p_3((N-3)\lambda + 3\mu) = p_2(N-2)\lambda + p_4 4\mu$$

$$\Leftrightarrow \quad p_4 = \frac{p_3((N-3)\lambda + 3\mu) - p_2(N-2)\lambda}{4\mu},$$

substituindo os valores das equações (4.38) e (4.39) obtemos:

$$p_4 = \frac{N(N-1)(N-2)(N-s+3)}{24} \left(\frac{\lambda}{\mu}\right)^4 p_0. \tag{4.40}$$

Note-se que $q_{k3} = 0$ se $\forall k \neq 2, 4$.

A entrada no estado 3 ocorre quando o processo sai do estado 2 para o estado 3 devido a chegada de um cliente a fila ou quando o processo sai do estado 4 para o estado 3 devido a saída de um cliente do sistema (o fluxo de entradas no estado 2 é assim quantificado por $(N-2)\lambda p_2 + 4\mu p_4$) e a saída do estado 3 ocorre quando o processo sai do estado 3 para o estado 4 devido a chegada de um cliente à fila ou quando o processo sai do estado 3 para o esta 2 devido a saída de um cliente do sistema (o fluxo de saídas do estado 1 é assim quantificado por $(N-3)\lambda p_3 + 3\mu p_3$.

Pelos resultados das distribuições estacionária obtidas nos estados anteriores podemos generalizar que:

$$p_n = \begin{cases} \frac{N!}{(N-n)!n!}(\frac{\lambda}{\mu})^n p_0, & se \quad 0 \leq n \leq s-1 \\ \frac{N!}{(N-n)!s!s^{n-s}}(\frac{\lambda}{\mu})^n p_0, & se \quad s \leq n \leq N \\ 0, & se \quad n > N \end{cases},$$

onde,

$$p_0 = \left[\sum_{n=0}^{s-1} \frac{N!}{(N-n)!n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s}^{N} \frac{N!}{(N-n)!s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1};$$

ainda assim, usando as equações (4.2) e (4.3), obtemos:

$$L_q = \sum_{n=s+1}^{N} (n-s)p_n;$$

$$L_s = \sum_{n=1}^{s-1} np_n + L_q + s\left(1 - \sum_{n=0}^{s-1} p_n\right);$$

daí,

$$W_s = \frac{L_s}{\overline{\lambda}} \qquad \text{e} \qquad W_q = \frac{L_q}{\overline{\lambda}},$$

onde $\overline{\lambda}$ calcula-se a partir da equação (4.36).

Tabela 4.6: Características do modelo $(M/M/s) : (FIFO/\infty/N)$

Chegada : Poissoneana Taxa : $\lambda_n = \begin{cases} (N-n)\lambda, & se \quad n = 0,1,2,...,N \\ 0, & se \quad n > N \end{cases}$	Tempo de atendimento: exponencial Taxa: $\mu_n = \begin{cases} n\mu, & se \quad n = 1,2,...,s-1 \\ s\mu, & se \quad n = s, s+1, ..., N \\ 0, & se \quad n > N \end{cases}$
População = N	Numero de servidores = s $(s > 1)$ Taxa de ocupação = $\frac{\overline{\lambda}}{s\mu}$ Taxa de desocupação = $1 - \frac{\overline{\lambda}}{s\mu}$
Número médio de clientes no sistema	$L_s = \sum_{n=1}^{s-1} np_n + L_q + s(1 - \sum_{n=0}^{s-1} p_n)$
Número médio de clientes que aguardam na fila	$L_q = \sum_{n=s+1}^{N} (n-s)p_n$
Tempo médio de espera de cliente no sistema	$W_s = \frac{L_s}{\overline{\lambda}}$
Tempo médio de espera de cliente na fila	$W_q = \frac{L_q}{\overline{\lambda}}$
Probabilidade de ocorrência do estado 0	$p_0 = \left[\sum_{n=0}^{s-1} \frac{N!}{(N-n)!n!}(\frac{\lambda}{\mu})^n + \sum_{n=s}^{N} \frac{N!}{(N-n)!s!s^{n-s}}(\frac{\lambda}{\mu})^n\right]^{-1}$
Probabilidade de ocorrência do estado n $p_n = \begin{cases} \frac{N!}{(N-n)!n!}(\frac{\lambda}{\mu})^n p_0, & se \quad 0 \le n \le s-1 \\ \frac{N!}{(N-n)!s!s^{n-s}}(\frac{\lambda}{\mu})^n p_0, & se \quad s \le n \le N \\ 0, & se \quad n > N, \end{cases}$	

4.6.7 Modelos de filas de espera não Markovianos

Nesta secção serão apresentados alguns modelos de filas de espera, em que os tempos entre chegadas de clientes e /ou os tempo de serviço não são exponenciais (modelos não-markovianos). Nestes casos, o raciocínio matemático é mais difícil mas, ainda assim, alguns autores apresentam resultados exatos ou aproximações úteis para várias características desses modelos. A analise matemática para estes modelos está para além deste trabalho, pelo que os resultados serão aqui apresentados de forma resumida.

4.6.7.1 Modelo $(M/G/1) : (FIFO/\infty/\infty)$

Este modelo assume que o processo de chegada de clientes é um processo de Poisson com intensidade (ou taxa) fixa λ, os tempos de serviço seguem um distribuição arbitrária (mas, a semelhança dos modelos anteriores, presume-se que os tempos são independentes) com valor médio $\frac{1}{\mu}$ e variância σ^2. Assume-se ainda que existe um único servidor, o atendimento é por ordem de chegada, a capacidade do sistema e a população são infinitos. A aplicaçaõ deste modelo pode-se ver em [5].

Para determinar várias características deste modelo, além do conhecimento do número de clientes no sistema, é preciso também saber quanto tempo o cliente demora para ser servido.

Suponhamos que é conhecido t_j, o instante de saída do j-ésimo cliente $j = 1, 2, ...$ Seja N_j o número de clientes no sistema no instante $t_j,\ j = 1, 2, ...$

Denotando por M_j a variável aleatória que representa o número de clientes que chegaram ao sistema durante o tempo de serviço do j-ésimo cliente, podemos escrever o seguinte:

$$N_{j+1} = N_j + (M_{j+1} - 1) + \delta,\ \ j = 1, 2, ..., \tag{4.41}$$

onde,

$$\delta = \begin{cases} 0, & se \quad N_j > 0 \\ 1, & se \quad N_j = 0 \end{cases}.$$

A partir da condição anterior verificam-se as seguintes identidades:

$$\delta^2 = \delta, \quad N_j\delta = 0, \quad \text{e} \quad N_j(1-\delta) = N_j. \tag{4.42}$$

Elevando ao quadrado ambos os membros da equação (4.41) e calculando o valor esperado tendo em conta as identidades (4.42), obtemos:

$$E[N_{j+1}^2] = E[N_j^2] + E[(M_{j+1} - 1)^2] + 2E[N_j]E[M_{j+1} - 1] + E[\delta]E[2M_{j+1} - 1],$$

uma vez as variáveis N_j e δ são independentes de M_{j+1}. Concluímos então que

$$E[N_j] = \frac{E[N_{j+1}^2] - E[N_j^2] - E[(M_{j+1} - 1)^2] - E[\delta]E[2M_{j+1} - 1]}{2E[M_{j+1} - 1]}.$$

No regime estacionário, tem-se:

$$E[N_{j+1}] = E[N_j], \qquad E[N_{j+1}^2] = E[N_j^2]$$

e, portanto

$$E[N_j] = \frac{E[(M_{j+1} - 1)^2] + E[\delta]E[2M_{j+1} - 1]}{2E[1 - M_{j+1}]}. \tag{4.43}$$

Usando a equação (4.41) temos que, no estado estacionário,

$$0 = E[M_{j+1}] - 1 + E[\delta] \Leftrightarrow E[\delta] = 1 - E[M_{j+1}].$$

Sabendo que o tempo de serviço de um cliente, T_s, é tal que $E[T_s] = \frac{1}{\mu}$ e $Var[T_s] = \sigma^2$ e usando a propriedade do valor médio condicionado temos:

$$E[M_{j+1}] = E\left[E[M_{j+1}|T_s]\right] = E[\lambda T_s] = \lambda E[T_s] = \frac{\lambda}{\mu} = \rho \tag{4.44}$$

e

$$\begin{aligned} E[M_{j+1}^2] &= E\left[E[M_{j+1}^2|T_s]\right] \\ &= E[(\lambda T_s)^2 + \lambda T_s] \\ &= \lambda^2 E[T_s^2] + \lambda E[T_s] \\ &= \lambda^2 \left(\sigma^2 + \frac{1}{\mu^2}\right) + \frac{\lambda}{\mu} \end{aligned}$$

$$= \lambda^2\sigma_s^2 + \rho^2 + \rho. \tag{4.45}$$

Substituindo (4.44) e (4.45) em (4.43), obtemos:

$$E[N_j] = \frac{\lambda^2\sigma^2 + \rho^2 + \rho - 2\rho + 1 + (1-\rho)(2\rho - 1)}{2(1-\rho)} = \rho + \frac{\rho^2 + \lambda^2\sigma^2}{2(1-\rho)}.$$

No estado estacionário, temos

$$L_s = E[N_j] = \rho + \frac{\rho^2 + \lambda^2\sigma^2}{2(1-\rho)},$$

conhecida como fórmula de Pollaczerk-Khintchine. Para obtermos os valores de W_s,W_q e L_q usa-se as equações (4.5), (4.6) e (4.7). Então,

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

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

e

$$L_q = \frac{\rho^2 + \lambda^2\sigma^2}{2(1-\rho)}. \tag{4.46}$$

Observação: Quando a distribuição de tempo de serviço é exponencial, $\sigma^2 = \frac{1}{\mu^2}$, e as características do modelo coincidem obviamente com os do modelo $(M/M/1) : (FIFO/\infty/\infty)$.

Tabela 4.7: Características do modelo $(M/G/1) : (FIFO/\infty/\infty)$

Taxa de ocupação	$\rho = \frac{\lambda}{\mu}$
Número médio de clientes no sistema	$L_s = \rho + \frac{\rho^2+\lambda^2\sigma^2}{2(1-\rho)}$
Número médio de clientes que aguardam na fila	$L_q = \frac{\rho^2+\lambda^2\sigma^2}{2(1-\rho)}$
Tempo médio de espera de cliente no sistema	$W_s = \frac{1}{\mu} + \frac{\rho^2+\lambda^2\sigma^2}{2\lambda(1-\rho)}$
Tempo médio de espera de cliente na fila	$W_q = \frac{\rho^2+\lambda^2\sigma^2}{2\lambda(1-\rho)}$
Probabilidade de ocorrência do estado 0	$p_0 = 1 - \rho$
Probabilidade de ocorrência do estado n	$p_n = \rho^n p_0$

4.6.7.2 Modelo $(G/G/1) : (FIFO/\infty/\infty)$

Este modelo assume que os tempos entre chegadas sucessivas de clientes e os tempos de serviços seguem distribuições arbitrárias, com média $\frac{1}{\lambda}$ e $\frac{1}{\mu}$, respetivamente, e que estes tempos são todos independentes. A disciplina de serviço é $FIFO$, existe um único servidor, a capacidade do sistema e a população são infinitas. Uma aplicação deste modelo pode-se consultar em [30].

Gross [17] mostrou que, para sistemas saturados (isto é, $1-\epsilon < \rho = \frac{\lambda}{mu} < 1$, para algum pequeno ϵ) é válida a seguinte aproximação para o tempo médio de espera de um cliente na fila:

$$W_q \approx \frac{\lambda(\sigma_a^2 + \sigma_s^2)}{2(1-\rho)},$$

onde σ_a^2 e σ_s^2 são as variâncias dos tempos entre chegadas sucessivas e dos tempos de serviço, respetivamente. Portanto, o número médio de clientes na fila é aproximadamente,

$$L_q = \lambda W_q \approx \frac{\lambda^2(\sigma_a^2 + \sigma_s^2)}{2(1-\rho)},$$

o tempo médio de espera de um cliente no sistema é aproximadamente

$$W_s = W_q + \frac{1}{\mu} \approx \frac{\lambda(\sigma_a^2 + \sigma_s^2)}{2(1-\rho)} + \frac{1}{\mu} = \frac{\lambda\mu(\sigma_a^2 + \sigma_s^2) + 2(1-\rho)}{2\mu(1-\rho)}$$

e o número médio de clientes na fila é aproximadamente

$$L_s = \lambda W_s \approx \frac{\lambda^2\mu(\sigma_a^2 + \sigma_s^2) + 2\lambda(1-\rho)}{2\mu(1-\rho)}.$$

4.6.8 Modelo de filas com prioridades

Os modelos de filas de espera com disciplina prioritária são os modelos em que a disciplina de fila é baseada num sistema de prioridades. Estes modelos assumem que existem K prioridades, onde 1 é considerada a mais alta prioridade e K a mais

Tabela 4.8: Características do modelo $(G/G/1) : (FIFO/\infty/\infty)$

Número médio de clientes no sistema	$L_s \approx \frac{\lambda^2\mu(\sigma_a^2+\sigma_s^2)+2\lambda(1-\rho)}{2\mu(1-\rho)}$
Número médio de clientes que aguardam na fila	$L_q \approx \frac{\lambda^2(\sigma_a^2+\sigma_s^2)}{2(1-\rho)}$
Tempo médio de espera de cliente no sistema	$W_s \approx \frac{\lambda\mu(\sigma_a^2+\sigma_s^2)+2(1-\rho)}{2\mu(1-\rho)}$
Tempo médio de espera de cliente na fila	$W_q \approx \frac{\lambda(\sigma_a^2+\sigma_s^2)}{2(1-\rho)}$
Taxa de ocupação	$\rho = \frac{\lambda}{\mu}$

baixa prioridade. Segundo [40] estes modelo também assume que os clientes são selecionados para o início do serviço na ordem da sua prioridade e, dentro de cada prioridade, os clientes são servidos por ordem de chegada, ou seja, a disciplina $FIFO$ é aplicada dentro de cada prioridade.

Os modelos de filas com disciplina prioritária são classificados em dois tipos:

1) Sistemas preemptivos, quando um cliente que está a ser servido não pode ser interrompido se um cliente de prioridade mais alta entrar no sistema, ou seja, qualquer cliente deve ser servido completamente sem interrupção. Como por exemplo, serviço no supermercados.

2) Sistemas não preemptivos, quando um cliente de prioridade baixa que está a ser servido é interrompido sempre que um cliente de prioridade mais alta entrar no sistema. O cliente cujo o serviço foi interrompido reentra e o serviço é retomado onde foi deixado. Como por exemplo, serviço na urgência de um hospital.

De seguida, apresentamos a notação para as filas com prioridades tem que ser adaptada e é a seguinte:

- λ_i = taxa de chegada para clientes da prioridade i, $(i = 1, 2, ..., K)$;
- $\lambda = \sum_{i=1}^{K} \lambda_i$ = taxa de chegada ao sistema;
- $\alpha_i = \frac{\lambda_i}{\lambda}$: percentagem de clientes de prioridade i que chegam ao sistema em uma determinada unidade de tempo;
- $E[S_i]$ = tempo médio de serviço para clientes que pertencem à prioridade i;
- $E[S_i^2]$ = segundo momento do tempo de serviço dos clientes que pertencem à prioridade i;
- $E[S] = \sum_{i=1}^{K} \alpha_i E[S_i]$: tempo médio de serviço;
- $\rho_i = \lambda_i E[S_i]$: fração de tempo que o servidor está ocupado com os clientes pertencentes à classe i;
- $\rho = \lambda E[S]$: taxa de ocupação (deve-se a segurar que $\rho < 1$ para que o sistema alcance um estado de equilíbrio);

- W_q^i = tempo médio gasto na fila para os clientes pertencentes à prioridade i;
- W_s^i = tempo médio no sistema para os clientes pertencentes à prioridade i;
- $W_s = \sum_{i=1}^{K} \alpha_i W_s^i$: tempo médio no sistema;
- L_q^i = número médio de clientes na fila que pertencem à prioridade i;
- L_s^i = número médio de clientes no sistema que pertencem à prioridade i;
- L_s = número médio de clientes no sistema;

Segundo [40] os modelos de filas com disciplina prioritária mantém que,

$$W_s^i = W_q^i + E[S_i] \tag{4.47}$$

$$L_q^i = \lambda_i W_q^i \tag{4.48}$$

$$L_s^i = \lambda_i W_s^i \tag{4.49}$$

$$L_s = \lambda W_s = \sum_{i=1}^{K} L_s^i. \tag{4.50}$$

Mais, W_q^i é dado por:

- Para modelos com prioridades preemptiva:

$$W_q^i = \frac{\sum_{j=1}^{K} \lambda_j E[S_j^2]}{2(1-\sigma_i)(1-\sigma_{i+1})}, \quad i = 1, 2, ..., K \tag{4.51}$$

- Para modelos com prioridades não-preemptivas:

$$W_q^i = \frac{E[S_i](1-\sigma_i) + \sum_{j=1}^{K} \frac{\lambda_j E[S_j^2]}{2}}{(1-\sigma_i)(1-\sigma_{i+1})} - E[S_i], \quad i = 1, 2, ..., K; \tag{4.52}$$

sendo $\sigma_i = \sum_{j=i}^{K} \rho_j$, $i = 1, 2, ..., K$ e $\sigma_{K+1} = 0$.

Capítulo 5

Aplicação da Teoria de Filas de Espera

5.1 Introdução

No Capítulo 4 foram estudados, sob o ponto de vista teórico, vários modelos de filas de espera e apresentaram-se, sempre que possível, várias características de cada um dos modelos. É através de tais características que se pode avaliar o funcionamento de um sistema de fila de espera e sugerir, quando for necessário e possível, modificações que melhorem o funcionamento do sistema. Tais modificações podem ser feitas nos elementos que definem o sistema, como por exemplo, o número de servidores, a capacidade do sistema, a disciplina de atendimento, etc..

Este capítulo é dedicado à apresentação e à exploração de exemplos de aplicações da teoria de filas de espera no mundo real. Existem aplicações em inúmeras áreas até porque, como já foi referido, enfrentamos filas de espera no nosso dia-a-dia nas mais variadas situações. No entanto, neste trabalho, dedicámo-nos a apresentar aplicações na área dos serviços de saúde (atendimento em hospitais, em clínicas de saúde, em laboratórios de análises clínicas, etc.). Tais aplicações podem incluir a análise de filas em quaisquer instalações de saúde e podem incidir sobre espaços, equipamentos e/ou pessoal [43]; em [12] e [14] podem-se consultar-se inúmeras referências a tais aplicações. Em algumas delas são usados modelos markovianos, que são os mais simples, mas noutras são usados modelos mais complexos (não markovianos e/ou com prioridades).

Nas duas secções seguintes são analisados com maior pormenor três artigos em que o uso de filas de espera foi relevante para a resolução de problemas em determinados serviços de saúde. No que resta desta secção vamos descrever, muito brevemente, outras situações reportadas na bibliografia e em que a teoria de filas de espera foi igualmente útil.

Preater [41] apresenta uma breve história do uso da teoria de filas de espera na área da saúde e aponta para uma extensa bibliografia que lista vários artigos. Apesar de não fornecer uma descrição detalhada das aplicações, é um artigo em que se pode perceber a importância desta teoria nesta área em particular. Segundo Vass e Szabo [51], a maioria dos artigos referindo aplicações de filas de espera na

área dos serviços de saúde foram publicados após 1990 e isso deveu-se ao avanço do poder computacional e à disponibilidade de software. Deste modo, esta teoria algo complexa passou a ser de mais fácil utilização para o pessoal médico e para os gestores de serviços de saúde.

Green [14] discute a relação entre tempos de espera longos e o número de servidores no modelo básico $(M/M/s) : (FIFO/\infty/\infty)$. O objetivo do estudo é determinar o número necessário de servidores de modo a reduzir a percentagem de pacientes que abandonam o hospital sem ser vistos devido ao tempo de espera longo.

Brahimi e Worthington [5] usam um modelo $(M/G/s) : (FIFO/\infty/\infty)$ para desenhar um sistema de marcação de consultas externas num hospital. Os autores constatavam que o sistema de marcação de consultas era frequentemente desenhado para minimizar o tempo de ócio do pessoal médico (tempo que um médico fica sem atender pacientes porque eles faltam ou chegam atrasados à consulta) o que tinha como consequência óbvia o aumento do tempo de espera dos pacientes. Neste trabalho, os autores propõe usar um sistema de chegadas que depende do instante de tempo t e desenhar o sistema de marcação de consultas de modo a incorporar a dependência de t. Deste modo, o sistema de fila de espera considerado foi, na verdade, um $(M(t)/G/s) : (FIFO/\infty/\infty)$ em que se usou uma cadeia de Markov não homogénea para o processo de chegadas.

Siddhartan, Jones e Johnson [46] consideram um modelo de fila de espera com prioridades (sendo usado o modelo markoviano com disciplina $FIFO$ dentro de cada prioridade) para analisar o funcionamento de uma urgência hospitalar. Em particular, os autores estudam os efeitos que a utilização da urgência para cuidados primários e/ou situações pouco graves tem sobre o tempo global de espera dos pacientes. Os autores concluem que a utilização de um sistema com prioridades diminui o tempo global de espera e diminui o tempo de espera dos pacientes mais graves (os de maior prioridade), mas pode fazer aumentar o tempo de espera dos pacientes de prioridades mais baixas.

5.2 Aplicação com modelos markovianos

Começamos por explorar um trabalho de Rosenquist [43], onde o autor usa o modelo $(M/M/1) : (FIFO/\infty/\infty)$ para averiguar como o aumento na taxa de chegada de pacientes afeta os tempos de espera e o comprimento da fila de espera num serviço de radiologia da urgência de um hospital. Usando conceitos básicos e um dos modelos de fila de espera mais simples, o autor apresenta um conjunto de conclusões importantes para o pessoal médico e os gestores hospitalares. Adicionalmente, o autor recomenda a utilização destes modelos a estes profissionais através de software específico para estes problemas.

O modelo de fila $(M/M/1) : (FIFO/\infty/\infty)$ usado em [43], assume que as chegadas dos pacientes ao serviço de radiologia se processam de acordo com um processo de Poisson, os tempos de serviço seguem uma distribuição exponencial, existe um único servidor, que neste caso é uma sala de raios-X, e os pacientes são servidos por sua ordem de chegada.

Para o estudo, as seguintes informações foram fornecidas pelo serviço de radiologia:

1) Os pacientes chegaram a uma taxa de 2 pacientes por hora, durante o primeiro ano de análise, e a taxa de chegada aumentou 5% em cada um dos 4 anos seguintes;

2) A taxa média de serviço manteve-se constante e igual a 3 pacientes por hora.

Convertendo estas informações em minutos, para o primeiro ano em análise, a taxa de chegada é igual a $\lambda = \frac{1}{30}$ pacientes por minuto, e a taxa média de serviço é igual a $\mu = \frac{1}{20}$ pacientes por minuto. Assim, a taxa de ocupação $\rho = \frac{\lambda}{\mu} = \frac{2}{3} = 0.(6)$, o que quer dizer que a utilização da unidade de raios-X é de 66.6%. Isto significa que, para o primeiro ano em análise, a máquina será usada em apenas dois terços do tempo. E, como $\rho < 1$, podem ser calculadas as seguintes características de interesse:

- a probabilidade de não haver nenhum paciente no sistema é

$$p_0 = 1 - \rho = 1 - 0.(6) = 0.(3);$$

- o número médio de pacientes na fila é

$$L_q = \frac{\rho^2}{1-\rho} = \frac{(\frac{2}{3})^2}{1-\frac{2}{3}} = 1.(3) \text{ pacientes};$$

- o tempo médio que um paciente aguarda na fila é

$$W_q = \frac{\rho}{\mu(1-\rho)} = \frac{\frac{2}{3}}{\frac{1}{20}(1-\frac{2}{3})} = 40 \text{ minutos};$$

- a probabilidade de um paciente ter que esperar é

$$P(N \geq 1) = \rho = 0.(6),$$

  isto é, dois terços dos pacientes têm que esperar para serem servidos.

Uma vez que a taxa de chegada teve um aumento de 5% em cada um dos 4 anos seguintes então, no quinto ano em análise, chegam

$$\lambda = 2 \times (1.05)^4 \simeq 2.43$$

pacientes por hora, ou $\lambda \simeq 0.04$ pacientes por minuto. As características acima referidas para o quinto ano em análise são alteradas para:

- taxa de ocupação:

$$\rho = \frac{\lambda}{\mu} \simeq \frac{0.04}{\frac{1}{20}} \simeq 0.81;$$

- a probabilidade de não haver nenhum paciente no sistema:

$$p_0 = 1 - \rho \simeq 1 - 0.81 = 0.19;$$

- o número médio de pacientes na fila:

$$L_q = \frac{\rho^2}{1-\rho} \simeq \frac{(0.81)^2}{1-0.81} \simeq 3.5 \text{ pacientes};$$

- o tempo médio que um paciente aguarda na fila:

$$W_q = \frac{\rho}{\mu(1-\rho)} \simeq \frac{0.81}{\frac{1}{20}(1-0.81)} \simeq 85.45 \text{ minutos};$$

- a probabilidade de um paciente ter que esperar:

$$P(N \geq 1) = \rho \simeq 0.81.$$

Os resultados da análise do serviço de radiologia ao longo dos 5 anos mostram que houve alterações drásticas nas principais características do sistema. Apesar de um aumento anual de apenas 5% nas chegadas dos pacientes, o tempo médio que um paciente aguarda na fila mais do que duplicou (passou de 40 para 85.45 minutos) e o número médio de pacientes na fila quase triplicou (aumentou de 1.(3) para 3.5 pacientes).

Ainda em [43], Rosenquist apresenta um outro modelo de fila de espera, o $(M/M/s) : (FIFO/k/\infty)$, com $1 < s < k$, para o serviço de radiologia em regime ambulatório, em que o número de servidores s e a capacidade do sistema k vão variar. Com este modelo, Rosenquist tem como objetivo estudar os efeitos provocados pelo abandono ou desistência de pacientes na fila de espera, uma vez que isto tem vários custos para a clínica. Note-se que, como este modelo corresponde a um sistema de capacidade limitada, haverá pacientes que não serão atendidos.

O modelo assume que as chegadas de pacientes se processam de acordo com um processo de Poisson, os tempos de serviço tem distribuição exponencial, existem s servidores, que são as salas de raios-X, e os pacientes são servidos por ordem de chegada. A sala de espera deste serviço de radiologia não comporta mais do que cinco pacientes, o que implica que, se um paciente chegar e encontrar cinco pacientes à espera de serem atendidos, ele não pode juntar-se a fila e o serviço de radiologia perde este paciente. Quando isto acontece o serviço de radiologia perde receitas, uma vez que não cobra o preço do serviço ao paciente que abandonou a clínica.

Neste trabalho, o objetivo de Rosenquist é precisamente estudar os custos que decorrem da perda de pacientes e, de uma forma geral, analisar o custo total do funcionamento do serviço de radiologia, incluindo o custo associado ao tempo de espera dos pacientes. Para esse efeito, o autor dispunha das seguintes informações:

- o custo do serviço de raios-X foi estimado em \$56 por hora, por paciente e por servidor;

- o custo do tempo que um paciente gasta no serviço de radiologia foi estimado em \$20 por hora;
- o custo de perder um paciente foi estimado em \$55 (isto representa o valor que a clínica perde por não cobrar o serviço ao paciente).

Considerando que os tempos de serviço dos diferentes servidores são independentes e identicamente distribuídos e tem uma média de 4 pacientes por hora ($\frac{1}{\mu} = 0.25$), o autor fez variar a taxa de chegada (λ) entre 6, 8 e 10 pacientes por hora, o número de servidores (s) entre 2 e 3, e a capacidade do sistema (k) entre 7 e 8. Para cada uma destas situações, Rosenquist determina algumas características do sistema como, por exemplo, o número médio de pacientes na fila, o tempo médio de espera de um paciente na fila, a probabilidade de um paciente ter que esperar na fila e a probabilidade de um paciente não ser servido, e estima o impacto do aumento das taxas de chegadas nos custos totais do serviço de radiologia.

Tabela 5.1: Análise do serviço de radiologia com chegadas de seis pacientes por hora

Número de servidores (s)	2	3
Taxa de chegada por hora (λ)	6	6
Taxa de serviço por servidor (μ)	4	4
Capacidade da sala de espera ($k - s$)	5	5
Capacidade do sistema (k)	7	8
Taxa de ocupação do serviço (%)	71.8	0.50
Probabilidade de um paciente ter que esperar	0.6	0.2
Probabilidade de um paciente não ser servido (p_7)	0.04	0.004
Taxa efetiva de chegada ($\overline{\lambda}$)	5.74	5.98
Tempo médio de espera de um pacientes no sistema (W_s)	0.43	0.29

A primeira situação a ser analisada corresponde ao modelo $(M/M/2) : (FIFO/7/\infty)$, com taxa de chegada de 6 pacientes por hora e os resultados encontram-se na segunda coluna da Tabela 5.1. Recorde-se que a taxa de serviço mantém-se sempre igual a 4 pacientes por hora. Para calcular o tempo médio de espera de um paciente no serviço de radiologia utilizou-se a equação (4.31). As probabilidades que constam dessa coluna foram obtidas do seguinte modo:

$$P(\text{"paciente ter que esperar"}) = 1 - (p_0 + p_1),$$

$$P(\text{"paciente não ser servido"}) = p_7,$$

em que p_0, p_1 e p_7 são obtidos de acordo com a Tabela 4.4. Com esta informação é possível concluir que este sistema custa ao serviço de radiologia \$175 por hora. De facto este custo é dado por:

$$\$56 \times 2 + \$20 \times W_s \times \overline{\lambda} + \$55\lambda \times p_7.$$

Na última coluna da Tabela 5.1 consta a informação para o caso em que existem 3 salas de raios-X e a taxa de chegadas de pacientes se mantém igual a 6. Fazendo os cálculos semelhantes aos acima indicados, verifica-se que o custo total do serviço

de radiologia aumenta para \$203. Assim, apesar da redução dos custos associados à diminuição do tempo de espera dos pacientes e à diminuição da probabilidade de um paciente não ser servido, o custo de uma máquina extra de raios-X não é compensado pela redução dos outros. Concluímos assim que, para uma taxa de chegadas de 6 pacientes por hora, o sistema ideal é aquele que tem apenas 2 máquinas de raios-X porque é o que tem custo total inferior.

De seguida, Rosenquist analisa o efeito que o aumento da taxa de chegada de pacientes e o aumento do número de salas de raio-X tem no custo total do funcionamento do serviço de radiologia. Assim, foram consideradas as seguintes situações:

- taxa de chegada de 8 pacientes por hora e 2 salas de raios-X;
- taxa de chegada de 8 pacientes por hora e 3 salas de raios-X;
- taxa de chegada de 10 pacientes por hora e 2 salas de raios-X;
- taxa de chegada de 10 pacientes por hora e 3 salas de raios-X.

Em todas estas situações, a taxa de serviço foi mantida constante igual a 4 pacientes por hora em cada uma das salas de raios-X e os resultados obtidos estão na Tabela 5.2.

Tabela 5.2: Análise do serviço de radiologia com chegadas de oito e dez pacientes por hora

	λ=8				λ=10			
Número de servidores	2		3		2		3	
Custo por servidor	\$56		\$56		\$56		\$56	
Custo de serviço		\$112		\$168		\$112		\$168
Custo do tempo no sistema/pp	\$20		\$20		\$20		\$20	
$\overline{\lambda}$	6.93		7.84		7.53		9.39	
W_s	0.52		0.33		0.63		0.38	
Custo de espera		\$71		\$51		\$95		\$71
Prob. de rejeição do serviço	0.12		0.02		0.25		0.06	
Custo por cliente perdido	\$55		\$55		\$55		\$55	
Custo do cliente perdido		\$52		\$9		\$136		\$34
Custo total		\$235		\$228		\$343		\$273

Da análise da Tabela 5.2, facilmente se verifica que, para as taxas de chegadas iguais a 8 ou 10, o sistema ideal é o que tem 3 salas de raios-X (isto é, o modelo $(M/M/3) : (FIFO/8/\infty)$), ao contrário do que aconteceu com taxa de chegadas igual a 6. Assim, quando as taxas de chegadas são superiores, apesar do custo que acarreta ter mais uma sala de raios-X, o custo total é menor devido às poupanças que se conseguem ter com a redução do tempo que um paciente gasta no serviço e com a redução da perda de pacientes.

Prosseguimos agora com a análise de um trabalho realizado por Khan e Callahan [22], onde os autores usam o modelo $(M/M/s) : (FIFO/\infty/\infty)$ para averiguar o

montante que um laboratório hospitalar deve investir em publicidade de modo a maximizar o lucro final. Neste trabalho, assume-se que o laboratório tem uma exigência de qualidade que consiste em que o tempo que um paciente aguarda na fila deve ser no máximo de 10 minutos (e, portanto, que o paciente abandona o sistema caso o seu tempo de espera exceda os 10 minutos) e pretende que esta exigência seja divulgada na publicidade com a esperança que isso faça aumentar o número de pacientes que se dirigem ao laboratório. Contudo, o aumento do número de pacientes deverá ser acompanhado do aumento do pessoal de modo a que tal exigência de qualidade se mantenha. Assim, o problema que é abordado neste trabalho é semelhante ao abordado por Rosenquist em [43] pois trata-se de perceber como se deve variar o número de servidores em função do aumento da taxa de chegadas, tentando minimizar os custos provocados por tempos de espera e a aquisição de material/pessoal.

Como ponto de partida, Khan e Callahan consideram que, no momento em que o estudo foi efetuado, o laboratório de análises clínicas funcionava de acordo com o modelo $(M/M/1) : (FIFO/\infty/\infty)$, com taxa de chegadas igual a 7 pacientes por hora, taxa de serviço igual a 9.23 pacientes por hora e um único servidor que é um técnico de análises clínicas (flebotomista). De seguida, analisam o efeito que o aumento das taxas de chegadas e o aumento do número de servidores tem nas principais características do modelo de filas de espera, com especial destaque para o tempo médio que um paciente espera na fila. Na Tabela 5.3 encontram-se os valores obtidos para várias características do sistema, tendo sido considerados os valores de $\lambda = 7, 15, 20, 25, 27$ pacientes por hora, o número de flebotomistas $s = 1, 2, 3, 4, 5$ e a taxa de serviço foi mantida constante igual a $\mu = 9.23$ pacientes por hora. De realçar que algumas combinações de λ, s e μ consideradas, resultam numa taxa de ocupação $\rho = \frac{\lambda}{s\mu} > 1$, pelo que neste casos o sistema não tem o estado estacionário e os resultados são omitidos na tabela. Os cálculos foram efetuados a partir da Tabela 4.1 (caso $s = 1$) e da Tabela 4.2 (caso $s > 1$).

Os resultados da Tabela 5.3 indicam que o sistema inicial tem um tempo médio de espera bastante alto (20.41 minutos). Para uma taxa de chegada de 15 pacientes por hora, o laboratório deve ter pelo menos 2 flebotomistas. No entanto, com apenas 2 flebotomistas, o tempo médio de espera na fila é demasiado alto (12.63 minutos) e decresce bastante se tivermos 3 ou mais flebotomistas. Para uma taxa de 20 pacientes por hora, menos de 3 flebotomistas não é adequado e com 3 flebotomistas o tempo médio de espera na fila já é bastante baixo (4.10 minutos). Finalmente, para as taxas de chegadas iguais a 25 ou 27 pacientes por hora, o sistema deve ter pelo menos 3 flebotomistas, mas são necessários 4 ou mais flebotomistas para que os tempos médios de espera na fila sejam inferiores a 10 minutos.

5.3 Aplicação com modelos não markovianos

Nesta secção será explorado um único artigo cientifico que envolveu a utilização de modelos de filas de espera em que os processos de chegadas e/ou tempos de serviço não correspondem a um M/M e envolveu ainda modelos com disciplina prioritária.

Tabela 5.3: Características do modelo $(M/M/s) : (FIFO/\infty/\infty)$, para diferentes valores de λ e s, e com $\mu = 9.23$ pacientes por hora

taxa de chegada	caraterística	número de pessoal				
		s=1	s=2	s=3	s=4	s=5
7	P_0	0.24	0.45	0.47	0.47	0.47
	L_s	3.14	0.89	0.77	0.76	0.76
	Lq	2.38	0.13	0.02	0.00	0.00
	Ws	26.91	7.59	6.63	6.52	6.50
	Wq	20.41	1.09	0.13	0.02	0.00
15	P_0	-	0.10	0.18	0.19	0.20
	L_s	-	4.78	1.96	1.69	1.64
	Lq	4.22	3.16	0.34	0.07	0.01
	Ws	-	19.13	7.84	6.76	6.55
	Wq	16.90	12.63	1.34	0.26	0.05
20	P_0	-	-	0.09	0.11	0.11
	L_s	1.86	-	3.53	2.42	2.23
	Lq	-	-	1.37	0.26	0.06
	Ws	-	-	10.60	7.27	6.68
	Wq	12.07	-	4.10	0.77	0.18
25	P_0	-	-	0.02	0.06	0.06
	L_s	1.59	-	10.35	3.53	2.91
	Lq	-	-	7.64	0.83	0.20
	Ws	-	-	24.84	8.48	6.98
	Wq	10.31	-	18.34	1.98	0.48
27	P_0	-	-	0.01	0.04	0.05
	L_s	-	-	40.23	4.23	3.23
	Lq	-	-	37.30	1.30	0.31
	Ws	-	-	89.40	9.39	7.18
	Wq	9.88	-	82.89	2.89	0.68

Lin, Patrick e Labeau [30], estudam, em conjunto, o acesso de pacientes à urgência (ED, do inglês "Emergency Department") e depois o seu seguimento, quando for o caso, para a unidade de internamento (IU, do inglês "Inpatient Unit"). Os autores consideram um modelo com duas filas de espera conectadas: uma primeira fila ascendente que modela o fluxo de pacientes que acede à ED e uma segunda fila a jusante que modela o acesso de uma parte destes pacientes que segue para a IU.

O fluxo de pacientes que acede à ED é modelado através de um modelo preemptivo $(M/G/c_1) : (FIFO/\infty/\infty)$ com várias prioridades e em que os servidores são as camas da ED. Já o fluxo de pacientes que segue da ED para a IU é modelado

através de um modelo $(G/G/c_2) : (FIFO/c_2/\infty)$ em que c_2 é o número de camas disponíveis na IU. Note-se que o acesso à ED é feito através de um sistema com prioridades (e dentro de cada prioridade a disciplina de atendimento é a $FIFO$), mas o acesso à IU não tem sistema de prioridades. Assim, se um paciente precisar de passar da ED para IU, ele só o pode fazer quando houver camas disponíveis na IU e fica a aguardar o tempo necessário na ED, ocupando aí uma cama. Com o modelo de duas filas conectadas, o objetivo dos autores era recalcular o tempo médio de atendimento de um paciente na ED, em função da quantidade de pacientes que ficam impedidos de seguir para IU, e usar esse valor para calcular o tempo médio de espera na fila para os pacientes das diferentes prioridades. É importante que estes últimos tempos estejam de acordo com normas superiormente impostas pelas autoridades. Obviamente, para atingir este objetivo foi necessário estudar o tempo médio que um paciente permanece na ED.

As cinco prioridades existentes no acesso à ED e os tempos médios de espera em cada uma delas, que são permitidos pelas autoridades competentes, são descritos na Tabela 5.4. O funcionamento conjunto da ED e da IU é descrito do seguinte modo: quando um paciente se desloca à ED, entra no sistema de triagem onde lhe vai ser atribuída uma das cinco prioridades. Ele aguarda um tempo, ocupando uma cama na ED, para ser visto por um médico que vai decidir por uma das seguintes situações:

1) o paciente recebe alta e abandona o ED, depois de cumpridas algumas formalidades, libertando assim uma cama;

2) o paciente precisa seguir para a IU. Neste caso, se não houver disponibilidades imediata de camas na IU, o paciente fica retido numa cama da ED e a aguardar vaga na IU.

Tabela 5.4: Niveis de prioridades e tempo de espera permitidos no DE

Niveis de prioridades	Tempo de espera para consultar um médico
I: Ressuscitação	Imediato
II: Emergente	< 15 minutos
III: Urgente	< 30 minutos
IV: Menos urgente	< 60 minutos
V: Não urgente	< 120 minutos

Para o estudo efetuado os autores tiveram acesso às taxas de chegadas das diferentes prioridades da ED e às taxas de chegadas da IU (via ED e também diretamente) do ano fiscal 2011/2012. Os valores destas taxas estão disponíveis na Tabela 5.5. Para além do conhecimento destas taxas, para estudar o tempo médio de espera até ao atendimento na ED, é necessário determinar o tempo médio de serviço da ED. De facto, uma vez que há pacientes que ficam a ocupar camas (servidores) na ED porque não tem camas disponíveis na IU, o tempo médio de serviço que seria usual

Tabela 5.5: Taxas de chegadas de pacientes à ED e à IU

Destino	Ponto de entrada	Taxas de chegada por hora
DE	I: Ressuscitação	0.075
	II: Emergente	0.662
	III: Urgente	3.749
	IV: Menos urgente	2.86
	V: Não urgente	0.226
	Todos	7.572
UI	a partir do DE	0.479
	Directo	0.267
	Todos	0.746

considerar para uma cama (servidor) da ED tem que ser recalculado para incorporar o tempo que os pacientes que aí ficam retidos.

Voltemos então ao modelo preemptivo $(M/G/c_1) : (FIFO/\infty/\infty)$, com cinco prioridades, que descreve a fila de espera da ED. Recordar que se assume que as chegadas de pacientes se processam de acordo com um processo de Poisson, que existem c_1 servidores (camas da ED) e que os tempos de serviço não têm distribuição especificada mas são independentes e identicamente distribuídos entre servidores. Se assumirmos que o tempo médio de serviço de cada um destes servidores é $\frac{1}{\mu}$ e se não houvesse retenção de pacientes na ED por falta de camas na IU, μ seria a usual taxa de serviço para esta fila. No entanto, neste caso, a taxa de serviço da ED será menor e o correspondente tempo médio de serviço de cada servidor da ED será dado por:

$$\frac{1}{\mu} = \frac{1}{\mu_1} + P_b \times E[\min\{T_1, T_2, ..., T_{c_2}\}], \tag{5.1}$$

onde P_b é a probabilidade de haver pacientes retidos na ED, T_i é o tempo que resta até ser dada alta ao paciente da i-ésima cama da IU $(i = 1, 2, .., c_2)$,

$$E[\min\{T_1, T_2, ..., T_{c_2}\}] = \int z dG(z)$$

com G a função de distribuição da variável aleatória $\min\{T_1, T_2, ..., T_{c_2}\}$. Recordar que em (2.1) está a expressão para se obter G. A taxa de pacientes transferidos da ED para IU será então dada por xR_d, com $R_d = \min\{\lambda, c_1\mu\}$, λ a taxa de chegadas de pacientes na ED e x a proporção de pacientes que necessita de transferência para IU.

Para calcular P_b é necessário recordar que esta representa a probabilidade de um paciente não conseguir aceder à IU. Uma vez que o modelo usado para a fila da IU é um $(G/G/c_2) : (FIFO/c_2/\infty)$, P_b é normalmente designada por probabilidade de bloqueio deste modelo e são conhecidas fórmulas para o seu cálculo fornecidas por Whitt [52]. Assim, se μ_I denotar a taxa de serviço de cada um dos servidores da IU (e é assumido que os tempos de serviço são independentes e identicamente

distribuídos entre servidores), então P_b é dada por

$$P_b = \frac{\alpha\beta e^{\frac{-k\beta}{v}}}{(1 - e^{\frac{-k\beta}{v}})\rho_d\sqrt{c_2}}, \tag{5.2}$$

sendo que: c_2 é o número de camas da UI, λ_d é a taxa de chegada dos pacientes que acedem diretamente a UI e os outros parâmetros são fornecidos pelas seguintes expressões

$$\rho_d = \frac{xR_d + \lambda_d}{c_2\mu_I},$$

$$\beta = \sqrt{c_2}(1 - \rho_d),$$

$$k = \sqrt{c_2},$$

$$v = \frac{1 + C_s^2}{2}$$

e

$$\alpha = \left[1 + \beta\frac{\Phi(\beta)}{\varphi(\beta)}\right]^{-1},$$

com C_s o coeficiente de variação do tempo de serviço de cada uma das camas da IU, $\Phi(\beta)$ e $\varphi(\beta)$ representam a função de distribuição e a função densidade de probabilidade de uma distribuição normal padrão, respetivamente.

Os autores deste artigo [30] discutem condições entre os valores de $\lambda, c_1, \mu, x, R_d$ e λ_d, c_2, μ_I para que o sistema inicial (formado pela ED e a IU) atinja o estado de equilíbrio. Se por um lado, temos que ter $\lambda \leq c_1\mu$ para que a primeira fila atinja o equilíbrio, a condição para a segunda fila é que $xR_d+\lambda_d \leq c_2\mu_I$. A conjugação destas duas condições tem um problema óbvio: para conhecer μ é necessário conhecer P_b que depende de R_d que, por sua vez, depende de μ. Não havendo uma expressão fechada para μ nem para P_b, os autores desenvolvem um método numérico para determinar μ.

Depois de terem encontrado uma forma de determinar μ, os autores prosseguem com o cálculo do tempo médio gasto por um paciente no sistema da ED (incluí o tempo de espera e o tempo de atendimento). Como este sistema tem cinco prioridades, é necessário começar por determinar este tempo para cada uma das prioridades. Em Bondi e Buzen [4], é fornecida a seguinte aproximação para o tempo médio gasto no sistema por um paciente da prioridade k ($k = 1, 2, ..., 5$):

$$W_{c_1}^k \approx \frac{W_{c_1}^F W_1^k}{W_1^F}, \tag{5.3}$$

em que W_1^k representa o tempo médio no sistema de um paciente da prioridade k numa fila preemptiva $(M/G/1) : (FIFO/\infty/\infty)$ o W_1^F representa o tempo médio no sistema de uma fila $(M/G/1) : (FIFO/\infty/\infty)$ e $W_{c_1}^F$ representa o tempo médio no sistema de uma fila $(M/G/c_1) : (FIFO/\infty/\infty)$. Deste modo, o valor de $W_{c_1}^k$ é obtido assim que se determinar os valores de W_1^k, W_1^F e $W_{c_1}^F$.

O valor de W_1^F é dado por $W_1^F = \frac{L_q}{\lambda}$, onde L_q é fornecido pela equação (4.46). Assim, tem-se

$$W_1^F = \frac{\rho^2 + \lambda^2\sigma^2}{2\lambda(1-\rho)} = \frac{\lambda^2(\sigma^2 + \frac{1}{\mu^2})}{2\lambda(1-\rho)} = \frac{\lambda(\sigma^2 + \frac{1}{\mu^2})}{2(1-\rho)} = \frac{\lambda E[S^2]}{2(1-\rho)}, \tag{5.4}$$

onde $\rho = \lambda E[S]$, λ representa a taxa de chegada de pacientes à ED, $E[S]$ representa o tempo médio de serviço da ED ($\frac{1}{\mu}$) e $E[S^2]$ representa o segundo momento do tempo de serviço da ED. Observar que $E[S^2]$ não é conhecido e vai ter que ser determinado.

Para determinar o valor de $W_{c_1}^F$, os autores remetem para Lee e Longton [29], que apresentam a seguinte aproximação

$$W_{c_1}^F \approx \left[\frac{1}{\mu} + W_q\right] \frac{1+C^2}{2}, \tag{5.5}$$

onde C representa o coeficiente de variação do tempo de serviço da ED e W_q, segundo Gross [17], é o tempo médio de espera na fila no modelo $(M/M/c_1) : (FIFO/\infty/\infty)$, isto é,

$$W_q = \frac{p_0(\frac{\lambda}{\mu})^{c_1}\rho}{\lambda c_1!(1-\rho)^2}, \quad (\rho = \frac{\lambda}{c_1\mu}).$$

Entao,

$$W_{c_1}^F \approx \left[\frac{1}{\mu} + \frac{p_0(\frac{\lambda}{\mu})^{c_1}\rho}{\lambda c_1!(1-\rho)^2}\right] \frac{1+C^2}{2}. \tag{5.6}$$

Podemos ainda escrever $W_{c_1}^F$ do seguinte modo:

$$W_{c_1}^F = \left[E[S] + \frac{P_Q E[S]}{c_1 - \lambda E[S]}\right] \frac{1+C^2}{2}, \tag{5.7}$$

onde

$$P_Q = \frac{(c_1\rho)^{c_1}}{c_1!(1-\rho)} \left[\sum_{t=1}^{c_1-1} \frac{(c_1\rho)^t}{t!} + \sum_{t=c_1}^{\infty} \frac{(c_1\rho)^t}{c_1! c_1^{t-c_1}}\right]^{-1},$$

$E[S]$ representa o tempo médio de serviço da ED e λ representa a taxa de chegada de pacientes à ED.

Finalmente, para calcular o W_1^k, usa-se a equação

$$W_1^k = W_q^k + \frac{1}{\mu},$$

onde W_q^k é obtido a partir da equação (4.51), ou seja,

$$W_1^k = \begin{cases} \frac{(1-\rho_1)E[S_1]+R_1}{1-\rho_1} & , \quad k=1 \\ \frac{(1-\rho_1-\ldots-\rho_k)E[S_k]+R_k}{(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1-\ldots-\rho_k)} & , \quad k>1, \end{cases} \tag{5.8}$$

onde $\rho_k = \lambda_k E[S_k]$, λ_k representa a taxa de chegada da prioridade k e $E[S_k]$ representa o tempo médio de serviço dos pacientes da prioridade k (depois de ter em conta os efeitos da conexão dos dois modelos), $R_k = \frac{1}{2}\sum_{i=1}^{k} \lambda_i E[S_k^2]$ e $E[S_k^2]$ representa o segundo momento do tempo médio de serviço dos pacientes da prioridade k.

Para terminar os cálculos, os autores desenvolveram ainda um algoritmo numérico que lhes permitiu obter $E[S^2]$ (necessário para o W_1^F) e ainda $E[S_k]$ e $E[S_k^2]$ necessários para o W_1^k.

Capítulo 6

Conclusões e Trabalhos Futuros

O trabalho apresentado nesta dissertação teve como principal objetivo explorar aplicações da teoria de filas de espera na área dos serviços de saúde.

Para atingir esse objetivo, o autor teve, em primeiro lugar, que adquirir conhecimentos de processos estocásticos, com particular destaque para as cadeias de Markov. Uma vez dominados esses conceitos, foi possível fazer um estudo aprofundado dos modelos de filas de espera mais conhecidos e mais utilizados na prática, com especial destaque para os modelos markovianos. No entanto, alguns modelos não-markovianos e/ou com disciplinas de atendimento prioritárias também foram ligeiramente abordados. Os modelos foram quase sempre estudados de modo a fornecer respostas a questões do tipo: quantos clientes aguardam em média na fila (e/ou no sistema), qual o tempo médio de espera de um cliente na fila (e/ou no sistema), qual a probabilidade de um cliente ter que esperar para ser atendido, etc., quando o processo se encontra no estado estacionário.

Depois de estudados os modelos de fila de espera mais importantes, foi feita uma pesquisa exaustiva na literatura sobre as aplicações de tais modelos aos serviços de saúde, uma vez que estes são frequentemente confrontados com problemas graves causados por atrasos nos atendimentos aos pacientes.

A partir desta pesquisa foi possível concluir que são realmente inúmeras as situações em que se utiliza teoria das filas de espera para tentar dar resposta a problemas relacionados com o congestionamento de serviços de saúde. Também foi possível observar a diversidade de investigadores (médicos, gestores, engenheiros, etc.) que se dedicam a estudar este tipo de problemas o que, por vezes, dificultou a compreensão da notação, da linguagem e da exposição dos conteúdos utilizada nos diversos artigos que foram explorados neste trabalho. De destacar também a grande diversidade dos modelos de filas de espera utilizados; desde o mais simples e clássico $(M/M/1) : (FIFO/\infty/\infty)$, a modelos com vários servidores e/ou capacidade finita e passando por modelos com disciplina de atendimento prioritárias, foi possível encontrar aplicações muito interessantes de todos eles e com resultados muito promissores.

Para trabalho futuro, recomenda-se explorar aplicações da teoria de filas de espera a outras áreas, como por exemplo, às ciências da computação, à economia e a telecomunicações. Do ponto de vista mais teórico, recomenda-se aprofundar o estudo de modelos não-markovianos, uma vez que vários autores têm obtido resultados e aproximações para as principais características de interesse destas filas.

Bibliografia

[1] Avi-Itzhak, B., Maxwell, W.L. and Miller, L.W. (1965) *Queuing with Alternating Priorities,* Operations Research, IS, No. 2 March-Apnl, pp. 306-318.

[2] Bailey, N.T. J. (1964) *The Elements of Stochastic Processes: With Applications to the Natural Sciences*, 1st edition, John Wiley Sons.

[3] Beveridge, R., Clarke, B., Janes, L. et al. (1998) *Implementation guidelines for the Canadian emergency department triage and acuity scale (CTAS).* Canadian association of emergency physicians.

[4] Bondi, A., Buzen, J. (1984) *The response times of priority classes under preemptive resume in M/G/m queues.* In ACM Sigmetrics, pp 195–201.

[5] Brahimi, M. and Worthington, D. J. (1991) *Queueing Models for Out-patient Appointment Systems – a Case Study.* The Journal of the Operational Research Society, 42, 733-746.

[6] Brumelle, S.L. (1971) *Some Inequalities for Parallel-Server Queues*, Operations Research 19, 402-413.

[7] Chan, T.C., Killeen, J.P., Kelly, D. et al. (2005) *Impact of rapid entry and accelerated care at triage on reducing emergency department wait times, length of stay, and rate of left without being seen.* Academic Emergency Medicine 46:491–497.

[8] Cooke, M.W., Wilson, S., Pearson, S. (2002) *The effect of a separate stream for minor injuries on accident and emergency department waiting times.* Emerg Med J 19(1):28–30. doi:10.1136/emj.19.1.28.

[9] Drake, A.W. (1967) *Fundamentals of Applied Probability Theory*, McGraw-Hill, New York.

[10] Feller, W. (1957) *An Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley.

[11] Fogliatti, M.C., Mattos, N.M.C. (2007) *TEORIA DE FILAS*, Brasil, Interciência.

[12] Fomundam, S., Herrmamnn, J. (2007) *A Survey of Queueing Theory Appliacations in Healthcare*, ISR TECHNICAL REPORT.

[13] Gonçalves, N., Lopes, N.M. (2012) *PROBABILIDADES Princípios Teóricos*, 2ª edição, Coimbra, Escolar Editora.

[14] Green, L.V. (2006a) *Queueing analysis in healthcare, in Patient Flow: Reducing Delay in Healthcare Delivery*, Hall, R. W., ed., Springer, New York, 281-308.

[15] Green, L.V. (2006b) *Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing*. Academic Emergency Medicine, 13, 61-68.

[16] Grimmett, G., Stirzaker, D. (2001) *Probability and Random Processes*. Oxford.

[17] Gross D, Shortle, J.F., Thompson, J.M. and Harris, C.M. (1985) *Fundamentals of queuing theory*, Wiley, New York.

[18] Hillier, F.S., Lieberman, G.J. (2006) *INTRODUÇÃO À PESQUISA OPERACIONAL*, 8ª Edição, McGraw-Hill.

[19] Hillier, F.S., Lieberman, G.J. (1995) *Introduction to Operations Research*, McGraw-Hill.

[20] Hokstad, P. (1978) *Approximations for the M/G/m queue*. Operations Research, 26:510–523.

[21] Karlin, S., Taylor, H.M. (1975) *A first course in stochastic processes*, Academic Press inc.

[22] Khan, M.R., Callahan, B.B. (1993) *Planning laboratory staffing with a queueing model*. European Journal of Operational Research, 67, 321-331.

[23] Kingman, J.F.C. (1962) *Some Inequalities for the Queue GI/G/1*, Biometrika 49, 315-324.

[24] Kingman, J.F.C. (1970) *Inequalities in the Theory of Queues*, Journal of Royal Statistical Society, Series B, v. 32, p. 120-110.

[25] Kleinrock, L. (1976) *Queueing Systems*, Vol. 1, Theory. Vol. 2, Computer Applications. New York: John Wiley & Sons, Interscience.

[26] Kolb, E.M.W, Taesik, L. et al (2007) *Effect of coupling between emergency department and IU on the overcrowding in emergency department*. Simulation Conference, 2007 Winter.

[27] Kumar, R., Sharma, S.K. (2012) *A Multi-Server Markovian Queueing System with Discouraged Arrivals and Retention of Reneged Customers*, International Journal of Operations Research, Vol. 9,No. 4, pp. 173-184.

[28] Kumar, R., Sharma, S.K. (2012) *M/M/1/N Queuing System with Retention of Reneged Customers*. Pak.j.stat. Operations Research. Vol. VIII No.4 pp. 859-866

[29] Lee, A.M., Longton, P.A. (1959) *Queueing process associated with airline passenger check-in*. Operations Research. Q 10:56–71.

[30] Lin, D., Patrick, J., Labeau, F. (2014) *Estimating the waiting time of multi-priority emergency patients with downstream blocking*, Health Care Management Sciences, 17:88-99.

[31] Marchall, W.G. (1978) *Some Simpler Bounds on the Mean Queuing Time*. Operations Research ,v. 16 n. 3, p. 651-665.

[32] Marchall, K.T. (1968) *Some inequalities in queuing*. Operations Research 16, 651–665.

[33] Molina, E.C. (1927) *Application of the Theory of Probability to Telefhone Trunking Problem*, Bell Syst. Tech.J.n. 6, p.461-494.

[34] Muller, D. (2011) *PROBABILIDADE E PROCESSOS ESTOCÁSTICOS*, Almedina. COIMBRA.

[35] Muller, D. (2007) *Processos Estocásticos e Aplicações*, Edições Almedina, SA.

[36] Murtela, B., Ribeiro, C. S., Silva, J. A., Pimenta, C. (2002) *INTRODUÇÃO À ESTATÍSTICA*, Portugal, MCGRAW-HILL.

[37] Nosek, Jr., R.A. and Wilson, J.P. (2001) *Queuing Theory and Customer Satisfaction: A Review of Terminology, Trends, and Applications to Pharmacy Practice*. Hospital Pharmacy, 36, 275- 279.

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

[39] Pestana, D., Velosa, S. (2010) *Introdução á Probabilidade e á Estatística*, Calouste Gulbenkian.

[40] Pardo, M.J, Fuente, D. (2007) *Optimizing a priority-discipline queueing model using fuzzy set theory*, Computers and Mathematics with Applications, 54, 267–281.

[41] Preater, J. (2002) *Queues in health.* Health Care Management Science, 5, 283.

[42] Rahman, K., Abdul, G.N., Kamil, A.A., Mustafa, A., Chowdhury, M.A.K. (2015) *An M/M/c/K State-Dependent Model for Pedestrian Flow Control and Design of Facilities*. PLoS ONE 10(7): e0133229.doi:10.1371/journal.pone.0133229.

[43] Rosenquist, C. J. (1987) *Queueing Analysis: A Useful Planning and Management Technique for Radiology*. Journal of Medical Systems, 11, 413-419.

[44] Ross, S.M. (2010) *Introduction to Probability Model*, 10th ed., Academic Press.

[45] Shimshak, D. G., Gropp Damico, D. and Burden, H.D. (1981) *A priority queuing model of a hospital pharmacy unit*. European Journal of Operational Research, 7, 350-354.

[46] Siddhartan K, Jones, W.J and Johnson, J.A. (1996) *A priority queuing model to reduce waiting times in emergency care*. International Journal of Health Care Quality Assurance, 9, 10-16.

[47] Suzuki, T. and Yoshida, Y. (1970) *INEQUALITIES FOR MANY-SERVER QUEUE AND OTHER QUEUES*, Journal of the Operations Reseach Society of Japan.

[48] Taylor, T.H., Jennings, A.M.C., Nightingale, D.A., Barber, B., Leivers, D., Styles, M. and Magner, J. (1969) *A study of anaesthetic emergency work. Paper 1: The method of study and introduction of queuing theory*. British Journal of Anaesthesia, 41, 70-75.

[49] Taylor, H. and Karlin, S. (1998) *An Introduction To Stochastic Modeling*, Academic Press.

[50] Tucker, J. B., Barone, J. E., Cecere, J., Blabey, R.G. and Rha, C. (1999) *Using queueing theory to determine operating room staffing needs*. Journal of Trauma, 46, 71-79.

[51] Vass, H., Szabo, Z. (2015) *Application of Queuing Model to Patient Flow in Emergency Department*. Case Study. Procedia Ecomomics and Finance, 32, 470-487.

[52] Whitt, W. (2004) *A Diffusion Approximation for the G/GI/n/m Queue*. Operations Research 52(6):922–941.

[53] Worthington, D. J. (1987) *Queueing Models for Hospital Waiting Lists*. The Journal of the Operation Research Society 38, 413-422.

[54] Worthington, D.J. (1991) *Hospital Waiting List Management Models*. The Journal of the Operational Research Society 42, 833-843.