Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Ednário Barbosa de Mendonça

Teoria de Filas Markovianas e Aplicações

Campina Grande
Julho de 2014

Ednário Barbosa de Mendonça

Teoria de Filas Markovianas e Aplicações

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientadora:
Divanilda Maia Esteves

Campina Grande
Julho de 2014

M539t Mendonça, Ednário Barbosa de.
Teoria de filas Markovianas e aplicações [manuscrito] / Ednário Barbosa de Mendonça. - 2014.
63 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Profa. Dra. Divanilda Maia Esteves, Departamento de Estatística".

1. Filas Markovianas. 2. Medidas de desempenho. 3. Variável aleatória. 4. Estatística. I. Título.

21. ed. CDD 519

Ednário Barbosa de Mendonça

Teoria de Filas Markovianas e Aplicações

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Aprovado em: / /

Banca Examinadora:

DMEsteves

Profª. Divanilda Maia Esteves
UEPB - DE/CCT
Orientadora

Gustavo H. E.

Prof. Gustavo Henrique Esteves
UEPB - DE/CCT
Examinador

Tiago Almeida de Oliveira

Prof. Tiago Almeida de Oliveira
UEPB - DE/CCT
Examinador

Dedicatória

Dedico a minha mãe, Ediane Alves Barbosa, por todo o apoio que ela tem me dado ao longo de toda uma vida e por todo o seu esforço em procurar fazer de mim uma pessoa melhor, e ao meu pai, José Nário Martins de Mendonça (*in memorian*).

Agradecimentos

A Deus pelas bençãos a mim concedidas e por nunca ter me desamparado ao longo de toda a minha vida.

A professora Divanilda Maia Esteves, por me orientar na construção deste trabalho, por me conceder a oportunidade de participar de um Projeto de Iniciação Científica e por ser sempre prestativa no decorrer da minha carreira acadêmica, tirando minhas dúvidas e fazendo com que eu buscasse sempre o aprendizado de forma correta, no intuito de enriquecer meu conhecimento, preparando-me assim para o futuro.

A família: minha mãe, que me ensinou o que é realmente importante na vida, que sempre me colocou em primeiro plano, que deixava de se presentear para me presentear em troca de um sorriso e um beijo e é a pessoa que mais amo nesse mundo, minha vó, Maria Odete Alves Barbosa, que está sempre ao meu lado, a minha noiva, Thuanne Barros de Oliveira, que me deu sempre força e apoio nos momentos difíceis, me entendia quando estava estressado e acreditou sempre em min quando ninguém mais acreditava, e aos meus tios, Hélio Alves Barbosa e José Israel Alves Barbosa, os quais também se fazem bastante presentes em minha vida.

Aos demais professores da UEPB que passaram por minha graduação e tiveram sua parcela de contribuição na construção da minha carreira acadêmica, em especial, os professores: Gustavo Henrique Esteves, Tiago Almeida de Oliveira, Ana Patrícia Barros Peixoto, João Gil de Luna, Ricardo Alves de Olinda, Onildo Freire, Vandik Estevam Barbosa e Francisco Guedes

Aos colegas e amigos do "Busão" da cidade de Cubati-PB, os quais passei toda minha graduação viajando até Campina Grande-PB e aos meus colegas de turma.

Resumo

Um sistema de filas pode ser definido como um sistema onde "usuários" chegam a um posto de atendimento, buscando algum serviço. O tempo entre chegadas é uma variável aleatória e o tempo gasto para realizar o serviço é outra variável aleatória. Devido a esse caráter aleatório é impossível garantir que términos de serviços coincidam exatamente com chegadas de usuários. Consequentemente há vezes que o serviço completa sua tarefa com um usuário e não encontra mais ninguém disponível com quem trabalhar, tornando assim, o sistema ocioso. Outras vezes um usuário chega e já encontra o serviço ocupado com alguma chegada anterior, então ele poderá aguardar a sua vez ou partir. Isso dependerá da estrutura do sistema, pois em uma fila de banco, por exemplo, o cliente pode esperar, mas quando se trata de uma ligação telefônica simples, em geral, não há opção de espera. Esses são aspectos básicos das filas, mas essas estruturas podem ser mais complexas, considerando outras situações como sistemas com uma capacidade finita de espera ou cliente desistindo do serviço quando demora a ser atendido. Atualmente, este tipo de estudo tem se destacado, especialmente devido às leis estabelecendo um tempo máximo de espera por atendimento em bancos, supermercados e *call centers*. O objetivo do estudo das filas é estimar os parâmetros envolvidos no modelo e calcular algumas medidas de seu desempenho, como por exemplo, tempo médio que o usuário fica na fila ou tamanho médio da fila, considerando as particularidades de cada caso. Uma vez que se conhece tais medidas, é possível buscar sistemas que atendam eficientemente às necessidades de quem procura o serviço sem que o sistema fique ocioso por muito tempo. Neste trabalho, aplicou-se a teoria das filas Markovianas ao fluxo de pessoas em uma casa lotérica da cidade de Cubati-PB, com o objetivo de comparar as medidas de desempenho em dias que há pagamento do benefício Bolsa Família do Governo Federal com os dias normais, ou seja, em que não há pagamento do benefício.

Palavras-chave: Filas Markovianas, Medidas de desempenho, Variável aleatória.

Abstract

A queuing system can be defined as a system where users get to a service station, looking for some service. The time between arrivals is a random variable and the time taken to perform the service is another random variable. Because of this randomness is impossible to guarantee that services endings coincide exactly with arrivals of users. Consequently there are times when the service completes its task with a user and can not find anyone else available to work with, thus making the system idle. Sometimes a user arrives and longer service busy with some earlier arrival then he can wait your turn or go. This depends on the structure of the system, as in a line at the bank, for example, the client can expect, but when it comes to a simple phone call, in general, no waiting option. These are basic aspects of the queues, but these structures may be more complex, considering other situations as systems with a finite capacity to hold or giving customer service takes when being serviced. Currently, this type of study has been outstanding, especially due to the laws establishing a maximum waiting time for service in banks, supermarkets and call centers. The objective of the study is to estimate the rows of the parameters involved in the model and calculate some measures of performance, such as average time that a user is in the queue or average queue size, considering the particularities of each case. Once you know these measures, it is possible to search efficiently systems that meet the needs of those looking for the service until the system is idle for long. This study applied the theory of Markovian flow of people queuing in a lottery town house Cubati-PB, with the aim of comparing the performance measures in days there Bolsa Família benefit payment from the Federal Government to normal days , meaning that no benefit payment.

Keywords: Markovian queues, performance measures, random variable.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

As filas de espera por serviços fazem parte do dia-a-dia das pessoas na sociedade e, como não podem ser evitadas, tendem a ser toleradas, apesar dos atrasos e das inconveniências que causam. Entretanto, os processos geradores de filas podem ser estudados e dimensionados de forma a aliviar os prejuízos em tempo e produtividade, assim como as perdas financeiras que elas acarretam. Entre as medidas que auxiliam no estudo de sistemas com fila, podem-se citar: número médio de elementos na fila, tempo de espera pelo atendimento e tempo ocioso dos prestadores de serviço.

Segundo Fogliatti e Mattos (2007), a Teoria de Filas consiste na modelagem analítica de processos ou sistemas que resultam em espera e tem como objetivo determinar e avaliar quantidades, denominadas medidas de desempenho, que expressam a produtividade e/ou operacionalidade desses processos. O estudo dessas quantidades é importante na tomada de decisão quanto à modificação ou manutenção da operação do sistema no seu estado atual, facilita também o dimensionamento racional da infraestrutura, de recursos humanos e financeiros, de equipamentos e instalações, visando um melhor desempenho no geral. Dessa forma, os conceitos e a teoria básica de Filas são fundamentais para a gerência e a administração de sistemas produtivos.

Em processos com determinadas características, após seu funcionamento durante um certo período de tempo, as medidas de desempenho tendem a se estabilizar. Neste caso, o intervalo de tempo de funcionamento do sistema, $[t_0, t)$, pode ser dividido em dois subintervalos: $[t_0, t^*)$ e $[t^*, t)$, onde t_0 é o instante de entrada em operação e t^* é o instante a partir do qual as medidas de desempenho se mantêm estáveis. Diz-se que em $[t_0, t^*)$ o sistema se encontra no regime transitório, enquanto que em $[t^*, t)$ o sistema se encontra no regime estacionário, os quais falaremos mais adiante.

No regime transitório, a variabilidade das medidas de desempenho dificulta as representações analíticas das mesmas, sendo necessários para tal, conhecimentos matemáticos avançados. No regime estacionário, a estabilidade dessas medidas permite o uso dos

respectivos valores esperados para a avaliação do sistema.

Neste trabalho, foi feito um estudo da Teoria das Filas Markovianas, bem como alguns conceitos prévios para a construção desta teoria. Alguns aspectos principais deste estudo serão apresentados na Fundamentação Teórica. Depois, aplicou-se tal teoria em um conjunto de dados referente ao fluxo de clientes em uma casa lotérica na cidade de Cubati-PB. O intuito foi comparar as medidas de desempenho para aqueles dias em que há pagamento do benefício Bolsa Família com aqueles dias em que não há.

2 Fundamentação Teórica

Boa parte deste trabalho consistiu em estudar a teoria envolvida no estudo das filas. Este capítulo contém alguns pontos importantes estudados. Primeiramente, serão apresentados os principais conceitos relacionados a Processos Estocásticos, Cadeias de Markov e Processos de Poisson, que são tópicos fundamentais para quem queira estudar Teoria de Filas. Em seguida, serão apresentados, de modo sucinto, algumas definições e resultados relacionados a essa teoria.

2.1 Processos Estocásticos

De modo geral, pode-se dizer que um processo estocástico é qualquer processo que evolui de maneira aleatória. Mais formalmente, segundo Fogliatti e Mattos (2007), um processo estocástico $\{X(t) : t \in T\}$ é uma coleção de variáveis aleatórias, isto é, para cada $t \in T$, $X(t)$ é uma variável aleatória. O conjunto T é chamado conjunto de índices. O conjunto de todos os valores que as variáveis $X(t)$ podem assumir é chamado espaço de estados S do processo estocástico.

Frequentemente, o índice t é interpretado como tempo t, e por isso, em geral, $X(t)$ é vista como o estado do processo no tempo t. Daí, de uma maneira alternativa, pode-se definir um processo estocástico como uma família de variáveis aleatórias que descreve a evolução de algum processo físico através do tempo.

Se $\{X(t), t \in T\}$ é um processo estocástico com espaço de estados S e conjunto de índices T, então considera-se que:

- Se S for enumerável, o processo é dito discreto ou a valores inteiros (do inglês, *integer-valued*). Se S é um intervalo da reta (ou o próprio $\mathbb{R}$) então dizemos que é um processo a valores reais (do inglês, *real-valued*)
- Se o conjunto de índices T for enumerável, então dizemos que o processo é a tempo

discreto e, em geral, consideramos $T = \{0, 1, 2, \dots\}$ e usamos $\{X_n, n \geq 0\}$ em lugar de $\{X(t), t \in T\}$. Se $T = [0, \infty)$, $X(t)$ é dito um processo a tempo contínuo.

Um aspecto importante que deve ser considerado é a estrutura de dependência associada à sequência de variáveis aleatórias. Em Estatística, um caso a se destacar é aquele em que as variáveis são independentes. Neste caso, o processo de inferência sobre os parâmetros envolvidos no modelo fica mais simples, uma vez que a distribuição conjunta das variáveis pode ser escrita como o produto das distribuições unidimensionais. Quando não se observa independência, busca-se descobrir o alcance e a forma de dependência. Um caso de dependência particularmente importante no estudo de processos estocásticos é a dependência de Markov.

Definição 2.1 *Um processo estocástico* $\{X(t), t \geq 0\}$ *é dito ser markoviano se para* $t_0 < t_1 < \cdots < t_{n+1}$,

$$P\left[X(t_{n+1}) = x_{n+1} | X(t_0) = x_0, X(t_1) = x_1, \dots, X(t_n) = x_n\right] = P\left[X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\right],$$

para qualquer escolha $x_0, x_1, \dots, x_{n+1}$ *em* S *e qualquer* n. *Isto quer dizer que, uma vez que conhecemos o estado atual da cadeia, os estados passados não influenciam o futuro.*

É possível generalizar um pouco mais esse caso, dizendo que um processo estocástico $\{X(t), t \geq 0\}$ *é markoviano de ordem* k *se para* $t_0 < t_1 < \cdots < t_{n+1}$,

$$P\left[X(t_n) = x_n | X(t_0) = x_0, X(t_1) = x_1, \dots, X(t_{n-1}) = x_{n-1}\right]$$

é igual a

$$P[X(t_n) = x_n | X(t_{n-k}) = x_{n-k}, \dots X(t_{n-1}) = x_{n-1}],$$

para qualquer escolha $x_0, x_1, \dots, x_{n+1}$ *em* S *e qualquer* n. *Isto quer dizer que o estado atual da cadeia é influenciado pelas* k *observações mais recentes do processo.*

Um processo Markoviano que possui o espaço de estados discreto é denominado Cadeia de Markov. O comportamento da Cadeia de Markov $\{X(t) : t \in T\}$ de parâmetro contínuo com espaço de estados discreto, o qual pode ser considerado sem perda de generalidade como $S = \{0, 1, 2, ...\}$, é caracterizado pela distribuição inicial

$$X(t_0) = l, \quad l = 0, 1, 2, ...,$$

onde t_0 é o instante inicial de observação, e pelas probabilidades (condicionais) de transição

entre os estados i e n, $P_{in}(v,z)$, definidas por:

$$P_{in}(v,z) = P[X(z) = n | X(v) = i], \quad 0 \leq v \leq z,\ v,z \in T,\ i,n \in S,$$

com:

$$P_{in}(v,v) = \begin{cases} 1, & \text{se } i = n, \\ 0, & \text{caso contrário} \end{cases}$$

2.2 Cadeias de Markov

Uma Cadeia de Markov é um Processo Estocástico com espaço de estados discreto. Se o conjunto de índices for um conjunto enumerável, então tem-se uma Cadeia de Markov a tempo discreto, caso contrário, tem-se uma Cadeia de Markov a tempo contínuo. Frequentemente, no caso de Cadeias de Markov, usa-se a notação $\{X_n, n \in \mathbb{N}\}$ em lugar de $\{X(t), t > 0\}$.

Definição 2.2 *Um processo $\{X_n : n > 0\}$ assumindo valores em um conjunto S é uma Cadeia de Markov se dado o estado presente, o futuro não é influenciado pelo passado, ou seja,*

$$P\left[X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, ..., X_n = x_n\right] = P[X_{n+1} = x_{n+1} | X_n = x_n].$$

Além disso, a Cadeia de Markov será dita homogênea no tempo se

$$P(X_{n+1} = y | X_n = x) = P(X_1 = y | X_0 = x) \tag{2.1}$$

para qualquer $n \geq 0$.

A probabilidade dada pela Equação (2.1) é chamada probabilidade de transição a um passo da cadeia e para facilitar a notação, usa-se

$$P(X_{n+1} = y | X_n = x) = P(X_1 = y | X_0 = x) = P(x,y),$$

ao que lê-se: a probabilidade de ir do estado x ao y em um passo.

Quando temos uma cadeia que é homogênea no tempo, suas probabilidades de transição podem ser representadas matricialmente na forma

$$P = \begin{bmatrix} P(0,0) & P(0,1) & P(0,2) & \cdots \\ P(1,0) & P(1,1) & P(1,2) & \cdots \\ P(2,0) & P(2,1) & P(2,2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Neste caso, P é a matriz de transição a um passo da cadeia, ou simplesmente matriz de transição da cadeia. Observe que os elementos da matriz são não negativos e que a soma dos elementos de cada linha deve ser igual a 1.

Basicamente, para que se possa modelar a evolução de uma Cadeia de Markov, é necessário conhecer como se comporta a cadeia inicialmente e como são feitas as transições a partir daí. As transições passo a passo são modeladas segundo a matriz de transições a um passo. No entanto, seria também interessante modelar as transições em um número maior de passos.

Definição 2.3 *A função* $\pi_0(x),\ x \in S$, *definida por*

$$\pi_0(x) = P(X_0 = x)$$

é chamada distribuição inicial da cadeia e é tal que

$$\pi_0(x) \geq 0$$

$$\sum_{x \in S} \pi_0(x) = 1$$

Frequentemente, a função de distribuição inicial é apresentada na forma de um vetor:

$$\pi_0 = [\pi_0(0), \pi_0(1), \pi_0(2), ...].$$

Teorema 2.1 *A distribuição conjunta de* $X_0, X_1, ..., X_n$ *pode ser expressa em termos da função de transição e da distribuição inicial da seguinte maneira:*

$$P(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = \pi_0(x_0)P(x_0, x_1)P(x_1, x_2)...P(x_{n-1}, x_n).$$

A ideia da demonstração do Teorema 2.1 é usar o Teorema da Multiplicação que pode ser encontrado na literatura clássica de probabilidade, que neste caso, seria

$$\begin{aligned} P(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = P(X_n = x_n | X_0 = x_0, X_1 = x_1, ..., X_{n-1} = x_{n-1}) \\ \times P(X_{n-1} = x_{n-1} | X_0 = x_0, X_1 = x_1, ..., X_{n-2} = x_{n-2}) \\ \times ... \times P(X_1 = x_1 | X_0 = x_0) P(X_0 = x_0). \end{aligned}$$

Definição 2.4 *A função de transição a m passos da cadeia é definida como*

$$P^m(x, y) = P(X_m = y | X_0 = x), \quad x, y \in S.$$

A função de transição a m passos da cadeia é a probabilidade de que uma cadeia, que está em um determinado estado x e em m "unidades de tempo", passe ao estado y em m passos, sem se importar com o que aconteceu "no meio do caminho". A função de transição a m passos também será representada na forma de uma matriz como

$$P^m = \begin{bmatrix} P^m(0,0) & P^m(0,1) & P^m(0,2) & \cdots \\ P^m(1,0) & P^m(1,1) & P^m(1,2) & \cdots \\ P^m(2,0) & P^m(2,1) & P^m(2,2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

2.2.1 Classificação dos Estados

Definição 2.5 *Um estado y é acessível a partir do estado x se $P_n(x, y) > 0$ para algum $n \geq 0$. Para dizer que y é acessível a partir de x, usa-se a notação $x \to y$.*

Note que isso implica que o estado y é acessível a partir do estado x se, começando em x, é possível que o processo, em algum tempo, atinja o estado y.

Definição 2.6 *Se dois estados x e y são acessíveis um a partir do outro, então x e y se comunicam e tal relação será denotada por $x \leftrightarrow y$.*

Observe que a relação de comunicação satisfaz as condições:

i) O estado x se comunica com ele mesmo, pois $P^0(x, x) = P(X_0 = x | X_0 = x) = 1$.

ii) Se o estado x se comunica com o estado y, isso implica que o estado y se comunica com o estado x.

iii) Se o estado x se comunica com o estado y e o estado y se comunica com o estado z, então o estado x se comunica com o estado z.

Definição 2.7 *Uma Cadeia de Markov é dita ser irredutível se existe apenas uma classe, isto é, se todos os estados se comunicam entre si.*

Definição 2.8 *Se x é tal que $P(x,x) = 1$, então x é um estado absorvente.*

Para cada estado $x \in S$, seja P_x a probabilidade de que, começando no estado x, o processo volte a atingir o estado x em algum tempo, isto é,

$$P_x = P(X_n = x, \text{para algum } n \geq 1 | X_0 = x).$$

Definição 2.9 *Um estado x é dito recorrente se $P_x = 1$. Se $P_x < 1$, então x é dito transitório.*

Proposição 2.1 *O estado x é recorrente se, e só se, $\sum_{n=1}^{\infty} P^n(x,x) = \infty$. Isto implica que, se $\sum_{n=1}^{\infty} P^n(x,x) < \infty$, então o estado é transitório.*

Corolário 2.1 *Se o estado x é recorrente e o estado y se comunica com o estado x, então o estado y é recorrente.*

Teorema 2.2 *Se $C \subset S$ é um conjunto finito fechado irredutível de estados, então todo estado de C é recorrente.*

Uma consequência dos resultados acima é que um estado transitório é visitado um número finito de vezes (daí o nome transitório). Isso implica que toda cadeia com espaço de estados finito deve ter pelo menos um estado recorrente. Além do mais, se a cadeia tiver espaço de estados finito e for irredutível, então todos os seus estados são recorrentes.

2.2.2 Medida Invariante

Definição 2.10 *Considere $\{X_n : n \geq 0\}$ uma Cadeia de Markov com espaço de estados S e função de transição P. Se $\pi(x),\ x \in S$ é tal que*

$$\pi(x) \geq 0,\ x \in S$$

$$\sum_{x \in S} \pi(x) = 1$$

além disso

$$\sum_{x \in S} \pi(x) P(x, y) = \pi(y) \tag{2.2}$$

então π é chamada medida invariante da cadeia.

A Equação (2.2) pode ser escrita na forma matricial como

$$\pi P = \pi,$$

onde P é a matriz de transição da cadeia e $\pi = (\pi(0), \pi(1), \pi(2), ...)$. Como consequência da definição, vemos que

- $\pi P = \pi$,
- $\pi P^2 = \pi P P = \pi P = \pi$,
- e de maneira geral $\pi P^n = \pi$, para todo $n \geq 1$.

Além do mais, se $\pi_0 = \pi$, então $\pi_n = \pi$ para qualquer $n \geq 1$.

2.3 Processo de Nascimento e Morte

De acordo com Fogliatti e Mattos (2007), uma Cadeia de Markov homogênea, irredutível, de parâmetro contínuo, é denominada Processo de Nascimento e Morte (*Birth-Death Process*) para todo $n > 0$ se,

$$\begin{aligned} P(n, n+1) &= \lambda_n \\ P(n, n-1) &= 1 - \lambda_n = \mu_n \\ P(n, k) &= 0, \quad \text{para qualquer } k \neq n-1, n+1. \end{aligned}$$

Em outras palavras, as únicas transições permitidas, em uma unidade de tempo, a partir de um determinado estado n são para seus vizinhos imediatos $n-1$ ou $n+1$. Quando a transição ocorre para estado $n+1$ representa um nascimento, e para o estado $n-1$, uma morte.

Para Processos de Nascimento e Morte, as seguintes hipóteses são válidas:

1. no instante inicial $t_0 = 0$, o sistema está vazio, isto é, $N(0) = 0$;

2. nascimentos e mortes são eventos estatisticamente independentes;

3. dado que o sistema está no estado n, no intervalo de tempo $(t, t+\Delta t)$, Δt tão pequeno quanto se queira, a probabilidade de ocorrer:

   (a) um nascimento é igual a $\lambda_n \Delta t + o(\Delta t)$;

   (b) uma morte é igual a $\mu_n \Delta t + o(\Delta t)$;

   (c) mais de um evento (nascimento(s) e/ou morte(s)) é desprezível, igual a $o(\Delta t)$, onde,

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

Usando a notação:

$$P_i^n(\Delta t) = P\{\text{ocorrência de } i \text{ nascimentos em } \Delta t\}$$

e

$$P_j^m(\Delta t) = P\{\text{ocorrência de } j \text{ mortes em } \Delta t\},$$

a terceira hipótese pode ser reescrita da forma a seguir:

$$P_1^n(\Delta t) = \lambda_n \Delta t + o(\Delta t), \quad \forall n = 0, 1, 2, ... \tag{2.3}$$

$$P_1^m(\Delta t) = \mu_n \Delta t + o(\Delta t), \quad \forall n = 1, 2, ... \tag{2.4}$$

$$P_i^n(\Delta t) P_j^m(\Delta t) = o(\Delta t), \quad \forall i, j | (i + j) > 1. \tag{2.5}$$

De (2.3), (2.4) e (2.5) obtêm-se as probabilidades de não haver nascimentos, Equação (2.6), nem mortes, Equação (2.7), em intervalos pequenos de tempo, Δt:

$$P_0^n(\Delta t) = 1 - \lambda_n \Delta t - o(\Delta t), \quad \forall n = 0, 1, 2, ... \tag{2.6}$$

$$P_0^m(\Delta t) = 1 - \mu_n \Delta t - o(\Delta t), \quad \forall n = 1, 2, ... \tag{2.7}$$

Com essas hipóteses, determinam-se, para todo n, as probabilidades P_n dos estados do processo, como mostrado a seguir.

Dividindo-se o intervalo de observação $(0, t + \Delta t)$ em dois sub-intervalos disjuntos $(0, t]$ e $(t, t + \Delta t)$, verifica-se que o sistema está no estado $n > 0$ no instante $(t + \Delta t)$, para Δt pequeno, se ocorre um dos seguintes eventos mutuamente excludentes:

1. no instante t, o sistema está no estado n e, no intervalo de tempo $(t, t + \Delta t)$, não há nenhum nascimento e nenhuma morte;

2. no instante t, o sistema está no estado $n-1$ e, no intervalo de tempo $(t, t+\Delta t)$, há um nascimento e não há nenhuma morte;

3. no instante t, o sistema está no estado $n+1$ e, no intervalo de tempo $(t, t+\Delta t)$, há uma morte e não há nenhum nascimento.

O sistema está no estado $n = 0$ no instante $(t + \Delta t)$, para Δt pequeno, se ocorre um dos seguintes eventos mutuamente excludentes:

1. no instante t o sistema está no estado 0 e, no intervalo de tempo $(t, t+\Delta t)$, não há nenhum nascimento;

2. no instante t o sistema está no estado 1 e, no intervalo de tempo $(t, t+\Delta t)$, há uma morte e nenhum nascimento.

Utilizando as hipóteses 1 e 2 acima e o Teorema da Probabilidade Total, pode-se calcular a probabilidade do sistema estar no estado n no instante $(t, t + \Delta t)$:

$$\begin{aligned} P_n(t+\Delta t) =& P_n(t)P_0^n(\Delta t)P_0^m(\Delta t) \\ &+ P_{n-1}(t)P_1^n(\Delta t)P_0^m(\Delta t) + P_{n+1}(t)P_0^n(\Delta t)P_1^m(\Delta t), \quad \forall n \geq 1, \end{aligned} \tag{2.8}$$

e

$$P_0(t+\Delta t) = P_1(t)P_0^n(\Delta t)P_1^m(\Delta t) + P_0(t)P_0^n(\Delta t) + o(\Delta t) \tag{2.9}$$

Substituindo as Equações (2.3), (2.4), (2.5), (2.6) e (2.7) em (2.8) e (2.9), têm-se:

$$\begin{aligned} P_n(t+\Delta t) &= P_n(t)[1-\lambda_n - o(\Delta t)][1-\mu_n\Delta t - o(\Delta t)] \\ &+ P_{n-1}(t)[\lambda_{n-1}\Delta t + o(\Delta t)][1-\mu_{n-1}\Delta t - o(\Delta t)] \\ &+ P_{n+1}(t)[1-\lambda_{n+1} - o(\Delta t)][\mu_{n+1}\Delta t - o(\Delta t)] + o(\Delta t) \\ &= P_n(t) - \lambda_n\Delta t P_n(t) - \mu_n\Delta t P_n(t) + Pn-1(t)\lambda_{n-1}\Delta t \\ &+ P_{n+1}(t)\mu_{n+1}\Delta t + o(\Delta t), \quad \forall n \geq 1, \end{aligned} \tag{2.10}$$

e ainda

$$\begin{aligned} P_0(t+\Delta t) &= P_0(t)[1-\lambda_0\Delta t - o(\Delta t)] + P_1(t)[1-\lambda_1\Delta t - o(\Delta t)][\mu_1\Delta t + o(\Delta t)] + o(\Delta t) \\ &= P_0(t) - \lambda_0\Delta t + \mu_1\Delta t P_1(t) + o(\Delta t), \end{aligned} \tag{2.11}$$

lembrando que $(\Delta t)^2 \cong o(\Delta t)$ e $o(\Delta t) \times \Delta t \cong o(\Delta t)$. De (2.10) e (2.11), obtêm-se:

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = -\lambda P_n(t) - \mu P_n(t) + \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) + \frac{o(\Delta t)}{\Delta t}, \quad \forall n \geq 1,$$

e

$$\frac{P_0(t+\Delta t)-P_0(t)}{\Delta t} = -\lambda_0 P_0(t) + \mu 1 P_1(t) + \frac{o(\Delta t)}{\Delta t}.$$

Tomando os limites quando $\Delta t \to 0$, têm-se:

$$\frac{dP_n(t)}{dt} = \lambda_n P_n(t) - \mu_n P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n-1} P_{n-1}(t), \forall n \geq 1, \tag{2.12}$$

$$\frac{dP_n(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t), \tag{2.13}$$

que formam um sistema infinito de equações diferenciais que representam as probabilidades dos estados do sistema.

Como o Processo de Nascimento e Morte é uma Cadeia de Markov irredutível, existe um tempo t^* a partir do qual ele entra no regime estacionário mantendo suas características estáveis. Neste caso,

$$\frac{dP_n(t)}{dt} = 0, \quad \forall n, \forall t > t^*.$$

Dessa forma, o sistema de equações diferenciais (2.12) e (2.13) se converte no sistema de equações algébricas

$$0 = -\lambda_n P_n - \mu_n P_n + \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1}, \quad \forall n \geq 1, \tag{2.14}$$

$$0 = -\lambda_0 P_0 + \mu_1 P_1. \tag{2.15}$$

Rearranjando (2.14) tem-se:

$$\lambda_n P_n - \mu_{n+1} P_{n+1} = \lambda_{n-1} P_{n-1} - \mu_n P_n \quad \forall n \geq 1$$

e usando recorrência,

$$\begin{aligned} \lambda_n P_n - \mu_{n+1} P_{n+1} &= \lambda_{n-1} P_{n-1} - \mu_n P_n \\ &= \lambda_{n-2} P_{n-2} - \mu_{n-1} P_{n-1} \\ &= \lambda_0 P_0 - \mu_1 P_1. \end{aligned} \tag{2.16}$$

De (2.15) e (2.16), tem-se:

$$\lambda_{n-1} P_{n-1} - \mu_n P_n = 0 \quad \forall n \geq 1.$$

Então,

$$P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \frac{\lambda_{n-1}\lambda_{n-2}}{\mu_n \mu_{n-1}} P_{n-2} = ... = \frac{\lambda_{n-1}\lambda_{n-2}\lambda_{n-3}...\lambda_0}{\mu_n \mu_{n-1}\mu_{n-2}...\mu_1} P_0$$

de onde se conclui que

$$P_n = P_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \quad n \geq 1. \tag{2.17}$$

Como $\sum_{n\geq 0} P_n = 1$, obtém-se:

$$P_0 = \frac{1}{1 + \sum_{n\geq 1} \prod_{i=1}^{n} \frac{\lambda_i - 1}{\mu_i}}. \tag{2.18}$$

desde que a soma do denominador de (2.18) seja convergente.

A partir das Equações (2.17) e (2.18), tem-se a distribuição limite dos estados do sistema $(P_0, P_1, P_2, ...)$, que é também a distribuição do estado do regime estacionário do processo, totalmente determinada pelas taxas de nascimento e morte.

As equações (2.14) e (2.15) são denominadas equações de balanço ou de equilíbrio, onde o princípio da conservação de energia é válido, ou seja, para cada estado, "o fluxo de quem entra é igual ao fluxo de quem sai".

Dessa forma, para qualquer estado $n \geq 1$, tem-se:

$$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = \lambda_n P_n.$$

e para cada estado $n = 0$,

$$\lambda_0 P_0 = \mu_1 P_1.$$

2.4 Processo de Poisson

Um Processo de Poisson ou de Nascimento Puro é uma Cadeia de Markov de parâmetro contínuo onde a única mudança permitida a partir de qualquer estado n é para o estado $n + 1$ e se processa com uma taxa constante. Então, esse processo pode ser modelado de forma análoga a um Processo de Nascimento e Morte, considerando-se as taxas de nascimento $\lambda_n = \lambda, \ \forall n$ e as taxas de morte $\mu_n = 0, \ \forall n \geq 1$.

Para o Processo de Poisson, as hipóteses consideradas para o Processo de Nascimento e Morte são válidas, a saber:

1. no instante inicial $t_0 = 0$, o sistema está vazio, isto é, $N(0) = 0$;

2. dado que o sistema está no estado n, no intervalo de tempo $(t, t + \Delta t)$, Δt tão

pequeno quanto se queira, a probabilidade de ocorrer:

(a) um nascimento é igual a $\lambda_n \Delta t + o(\Delta t)$;

(b) mais de um evento (nascimento(s)) é desprezível, igual a $o(\Delta t)$, onde,

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

As hipóteses 3a) e 3c) acima podem ser reescritas como,

$$P_1(\Delta t) = \lambda \Delta t + o(\Delta t), \tag{2.19}$$

$$P_i(\Delta t) = o(\Delta t), \quad \forall i \geq 1. \tag{2.20}$$

A partir das hipóteses 1 e 2, de (2.19), de (2.20) e utilizando-se o Teorema da Probabilidade Total, têm-se:

$$\begin{aligned} P_n(t + \Delta t) &= P_n(t)P_0(\Delta t) + P_{n-1}(t)P_1(\Delta t) + o(\Delta t) \\ &= (1 - \lambda \Delta t)P_n(t) + \lambda \Delta t P_{n-1}(t) + o(\Delta t), \quad n \geq 1, \end{aligned}$$

o que implica que

$$P_n(t + \Delta t) = P_0(t)P_0(\Delta t) + o(\Delta t) = P_0(t)(1 - \lambda \Delta t) + o(\Delta t).$$

Usando um procedimento análogo àquele usado para Processos de Nascimento e Morte, pode-se escrever:

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(\Delta t)}{\Delta t}, \quad \forall n \geq 1$$

e

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -\lambda P_0(t) + \frac{o(\Delta t)}{\Delta t}.$$

Tomando-se os limites quando $\Delta t \to 0$, têm-se:

$$\frac{dP_n(t)}{dt} = -\lambda[P_n(t) - P_{n-1}(t)], \quad \forall n \geq 1,$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t), \tag{2.21}$$

que constituem um sistema infinito de equações diferenciais. A solução da Equação (2.21) é dada por:

$$P_0(t) = e^{-\lambda t}. \tag{2.22}$$

Resolvendo-se (2.21) recorrentemente, têm-se:

$$\begin{aligned} P_1(t) &= \lambda t e^{-\lambda t}, \\ P_2(t) &= \frac{(\lambda t)^2 e^{-\lambda t}}{2!}, \\ P_3(t) &= \frac{(\lambda t)^3 e^{-\lambda t}}{3!}, \\ &\vdots \end{aligned}$$

de onde, por indução,

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \quad \forall n \geq 0.$$

Observa-se que a variável aleatória discreta que representa o estado do processo em um intervalo de comprimento t segue uma distribuição de Poisson de parâmetro λt, com valor esperado dado por:

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

O Processo de Poisson pode ser descrito de forma equivalente pela caracterização do tempo entre mudanças sucessivas como uma variável aleatória exponencial: seja T a variável aleatória que representa o tempo entre ocorrências sucessivas em um Processo de Poisson de taxa λ, e seja $N(t)$ a variável aleatória que representa o número de ocorrências num intervalo de tempo de comprimento t.

O evento $(T > t)$ é equivalente ao evento $(N(t) = 0)$, então:

$$P(T > t) = P(N(t) = 0) = P_0(t). \tag{2.23}$$

De (2.22) e (2.23), pode-se determinar a função de distribuição acumulada de T,

$$F_T(t) = 1 - P(T > t) = 1 - P_0(t) = 1 - e^{-\lambda t}, \quad t \geq 0$$

e a função de densidade de probabilidade,

$$f(t) = \frac{dF(t)}{dt} = -\lambda e^{-\lambda t} \quad t \geq 0$$

que é a função de densidade de uma variável aleatória com distribuição exponencial de parâmetro λ.

Reciprocamente, se os tempos T entre mudanças sucessivas são variáveis aleatórias independentes, idênticas e exponencialmente distribuídas de parâmetro λ, a variável

aleatória discreta $N(t)$ que representa o estado do sistema no tempo t segue uma distribuição de Poisson de parâmetro λt.

Para demonstrar essa propriedade, seja T_{n+1} a variável aleatória que representa a soma dos tempos transcorridos entre $n+1$ mudanças sucessivas. O evento $(T_{n+1} > t)$ é equivalente ao evento $(N(t) \leq n)$, então,

$$P(T_{n+1} > t) = P(N(t) \leq n) = \sum_{i=0}^{n} P_i(t) = F_N(n) \tag{2.24}$$

onde F_N é a função de distribuição acumulada de $N(t)$.

Como os tempos entre mudanças sucessivas seguem idênticas distribuições exponenciais de parâmetro λ, T_{n+1} tem distribuição de Erlang de parâmetros $(n+1)$ e λ. Dessa forma,

$$P(T_{n+1} > t) = \int_t^{\infty} \frac{\lambda(\lambda x)^n}{n!} e^{-\lambda x} dx$$

Fazendo-se a transformação $u = x - t$ e observando-se que t é uma constante, obtém-se:

$$\begin{aligned}
P\left(T_{n+1} \geq t\right) &= \int_t^{\infty} \frac{\lambda^{n+1}(u+t)^n}{n!} e^{-\lambda u} e^{-\lambda t} du \\
&= \int_t^{\infty} \frac{\lambda^{n+1} e^{-\lambda u} e^{-\lambda t} \sum_{i=0}^{n} \binom{n}{i} u^{n-i} t^i}{n!} du \\
&= \int_t^{\infty} \frac{\lambda^{n+1} e^{-\lambda u} e^{-\lambda t} \sum_{i=0}^{n} \frac{n!}{(n-i)!i!} u^{n-i} t^i}{n!} du \\
&= \sum_{i=0}^{n} \frac{\lambda^{n+1} e^{-\lambda t} t^i}{(n-i)!i!} \int_t^{\infty} e^{-\lambda u} u^{n-i} du \\
&= \sum_{i=0}^{n} \frac{\lambda^{n+1} e^{-\lambda t} t^i}{(n-i)!i!} \cdot \frac{\int_t^{\infty} e^{-\lambda u} (\lambda u)^{n-i}}{\lambda^{n-i+1}} \lambda du \\
&= \sum_{i=0}^{n} \frac{\lambda^{n+1} e^{-\lambda t} t^i}{(n-i)!i!} \cdot \frac{\Gamma(n-i+1)}{\lambda^{n-i+1}}
\end{aligned}$$

onde $\Gamma(.)$ é a função Gama, definida por

$$\Gamma(p) = \int_0^{\infty} e^{-u} u^{p-1} du.$$

Para p inteiro,

$$\Gamma(p) = (p-1)!$$

Dessa forma,

$$P(T_{n+1} > t) = \sum_{i=0}^{n} \frac{\lambda^{n+1} e^{-\lambda t} t^i (n-i)!}{(n-i)! i! \lambda^{n-i+1}} = \sum_{i=0}^{n} \frac{(\lambda t)^i e^{-\lambda t}}{i!} \tag{2.25}$$

De (2.24) e (2.25)

$$F_N(n) = \sum_{i=0}^{n} \frac{(\lambda t)^i e^{-\lambda t}}{i!}$$

o que caracteriza $N(t)$ como uma variável de Poisson de parâmetro λt.

Uma propriedade da distribuição exponencial é a "ausência de memória", denominada também propriedade Markoviana, que garante que dada uma informação presente sobre uma variável aleatória Exponencial, seu comportamento futuro independe do passado, isto é,

$$P(T \leq t_1 | T \geq t_0) = P(0 \leq T \leq t_1 - t_0).$$

A variável aleatória exponencial é a única variável aleatória contínua com essa propriedade.

2.5 Conceitos Básicos da Teoria de Filas

Segundo Hillier e Lieberman (1974), a Teoria de Filas estuda a espera em diversas formas de filas. Tal teoria busca definir maneiras de lidar mais eficientemente com sistemas de filas. Esta definição é possível a partir da determinação da maneira como o sistema de filas funcionará e o tempo médio de espera nas mesmas a partir da utilização de modelos de filas para diversas situações reais. Fogliatti e Mattos (2007) definem um Sistema com Fila como qualquer processo onde usuários oriundos de uma determinada população chegam para receber um serviço pelo qual esperam, se for necessário, saindo do sistema assim que o serviço é completado. Essa espera acontece quando a demanda é maior do que a capacidade de atendimento oferecida, em termos de fluxo.

O estudo da Teoria das Filas teve início com o matemático A.K. Erlang no ano de 1909 para o problema de congestionamento de linhas telefônicas na Dinamarca. A.K. Erlang é considerado por alguns autores o "pai" da Teoria de Filas, devido ao fato de seu trabalho ter se antecipado por várias décadas aos conceitos modernos dessa teoria. Já no ano de 1917, publicou o livro "*Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges*", onde sua experiência ficou documentada.

Desde então, as áreas de economia, administração e de processamento de fluxos usu-

fruíram dessa técnica, destacando-se, entre outros, problemas de congestionamento de tráfego, escoamento de fluxo de carga de terminais, carregamento/descarregamento de veículos, escoamento e fluxo de processamento de informações, formação de estoques, comunicação de computadores.

Dentre os trabalhos precursores desenvolvidos em diferentes áreas, podem-se citar, as modelagens apresentadas nas referências: Adams (1936) e Tanner (1951), para o cálculo do tempo médio de espera de pedestres para atravessar uma rua sem sinal; Everett (1953), para o problema de escoamento de fluxo de barcos em terminais portuários; Cobham (1954), para o problema de reparo de maquinárias; Bailey (1954), para o problema de escoamento do fluxo de pacientes na emergência de um hospital; Morse (1962) e Prabhu (1965), para o problema de formação de estoques; Bitran e Morabito (1996), para sistemas de manufaturas.

A partir de 1960, a Teoria de Filas foi também utilizada para modelar problemas concernentes à Ciência da Computação, Ramamoorthy (1965) e Courtois (1977). Destacam-se a partir de 1980, as aplicações a redes de filas, Gelembe e Pujolle (1987) e Walrand (1988); em comunicação de computadores, Daigle (1991) e em provedores de internet, Fontanella e Morabito (2001).

Das publicações em português sobre Teoria de Filas, podem-se mencionar Novaes (1975), que apresenta aplicações direcionadas ao Planejamento de Transportes, e MAGALHÃES (1996), que trata de modelos de redes de filas frequentemente aplicados à Ciência da Computação.

2.5.1 Estrutura Básica de um Sistema com Fila

Um sistema com fila é composto por usuários, por canais ou postos de serviço/atendimento e por um espaço designado para a espera.

Os usuários chegam segundo um determinado comportamento que caracteriza o processo de chegadas, para serem atendidos em canais ou postos de serviço (que funcionam em paralelo) segundo um padrão de atendimento. Enquanto os postos estão ocupados, os usuários aguardam numa única fila em um espaço designado para tal. Assim que um canal de serviço fica livre, um dos usuários da fila é chamado para atendimento segundo um critério estabelecido pela gerência. Uma vez completado o serviço, o usuário é liberado do sistema. A Figura 1 representa esquematicamente um sistema com fila.

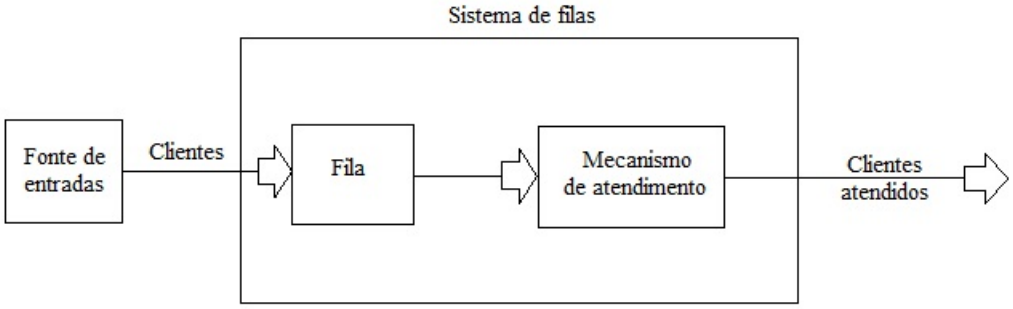

Figura 1: Representação esquemática de um sistema com fila

O processo de chegadas dos usuários é especificado pelo comportamento do fluxo de chegadas dos mesmos ao sistema. Segundo Fogliatti e Mattos (2007), se são conhecidos o número de chegadas e os instantes de tempo em que elas acontecem, esse processo é denominado determinístico, caso contrário, tem-se um comportamento aleatório constituindo um processo estocástico caracterizado por uma distribuição de probabilidade. O caso mais comum e simples, é quando se considera que os clientes chegam segundo um processo de Poisson.

O processo de atendimento é especificado pelo comportamento do fluxo de usuários e a sua caracterização é análoga à do processo de chegadas.

Os canais ou postos de serviço são os locais onde são atendidos os usuários. O número de postos de um sistema pode ser finito ou infinito. Como exemplo do primeiro caso, podem-se citar os guichês de um posto de pedágio, para o segundo caso, qualquer atendimento do tipo *self-service*, onde cliente e servidor são a mesma pessoa e onde o serviço está sempre disponível.

Segundo Fogliatti e Mattos (2007), a capacidade do sistema é o número máximo de usuários que o mesmo comporta (incluindo fila e atendimento) e pode ser finita ou infinita. Quando a capacidade é finita, os clientes que chegam ao sistema após a capacidade máxima ser atingida são rejeitados. Para o caso de capacidade infinita pode-se citar a espera de navios em ambiente aquaviário para descarregamento em um porto.

2.5.2 Disciplina de Atendimento

A disciplina de atendimento consiste na maneira pela qual os usuários que estão na fila são selecionados para serem atendidos. Os tipos de disciplinas de atendimento mais utilizados são:

- $FIFO$ (*first in - first out*): os usuários são atendidos na ordem das chegadas. Essa disciplina de atendimento é a mais comumente adotada.
- $LIFO$ (*last in - first out*): o primeiro usuário a ser atendido é o que chegou por último.
- PRI (*priority service*): o atendimento aos usuários segue uma ou mais prioridades preestabelecidas pela gerência do sistema.
- $SIRO$ (*service in random order*): o atendimento aos usuários segue uma ordem aleatória.

Vale salientar que há outros tipos de disciplinas de atendimento, inclusive considerando aspectos como atendimento prioritário e desistências. No entanto, como este é apenas um estudo introdutório, tais modelos não foram vistos.

2.5.3 Notação de um Sistema com Fila

A notação utilizada neste trabalho para descrever um sistema com fila é a notação encontrada em boa parte da literatura clássica de estudo de filas e foi proposta por Kendall (1953). Considera-se a forma $A/B/C/D/E$, onde A e B denotam, respectivamente, as distribuições dos tempos entre chegadas sucessivas e de atendimento, C e D denotam o número de postos de atendimento em paralelo e a capacidade física do sistema, respectivamente e E, uma das siglas que representam as disciplinas de atendimento.

Como exemplos de escolhas para A e B, podem-se citar:

- D: distribuição determinística ou degenerada; e para comportamento aleatório;
- M: distribuição exponencial (*Memoryless ou Markoviana*);
- E_k: distribuição Erlang do tipo k;
- G: distribuição geral (não especificada).

Para simplificar a notação, frequentemente as letras D e E da notação acima descrita são omitidas. Quando tais "parâmetros" não aparecem, considera-se que o sistema tenha capacidade infinita e disciplina de atendimento $FIFO$.

2.5.4 Medidas de Desempenho

Segundo Fogliatti e Mattos (2007), a utilização da Teoria de Filas permite avaliar a eficiência de um sistema por meio da análise de suas características utilizando medidas de operacionalidade/desempenho. Essas características, na maioria das vezes, mudam ao longo do tempo, sendo então representadas por variáveis aleatórias, cujos valores esperados podem ser utilizados como medidas de desempenho do sistema no regime estacionário. Dentre essas medidas, podem-se citar:

- Número médio de usuários na fila (L_q) e no sistema (L).
- Tempo médio de espera de um usuário qualquer na fila (W_q).
- Tempo médio de permanência de um usuário qualquer no sistema (W).

Outras medidas de desempenho que caracterizam o comportamento do sistema são:

- Probabilidade de se ter no máximo um número n_0 pré-fixado de usuários no sistema, $P(N \leq n_0)$.
- Probabilidade de um usuário qualquer ter que aguardar mais do que um determinado tempo t na fila, $P(T_q > t)$.
- Probabilidade de se ter algum servidor ocioso em um sistema com c postos de atendimento, $P(N < c)$.

2.6 Modelos de Filas Básicos

Nesta seção, serão apresentados modelos de filas representando situações que se comportam como processos Markovianos de "nascimento e morte".

2.6.1 Modelo $M/M/1/\infty/FIFO$

O modelo $M/M/1/\infty/FIFO$, como foi dito anteriormente, pode ser representado como $M/M/1$ para simplificar a notação. Este modelo é caracterizado por:

- tempos entre chegadas sucessivas e os tempos de atendimento seguindo distribuições exponenciais;
- existe um único posto de atendimento;
- não há limitação para o espaço reservado para a fila de espera;
- a ordem de acesso de usuários ao serviço segue a ordem das chegadas dos mesmos ao sistema.

As chegadas e os atendimentos caracterizam um Processo de Nascimento e Morte, lembrando que somente um único evento pode acontecer em períodos pequenos de tempo. As taxas de chegada (ingresso) ao sistema e de atendimento são constantes e dadas respectivamente por:

$$\lambda_n = \lambda \quad \forall n \geq 0$$

e

$$\mu_n = \mu \quad \forall n \geq 1.$$

Segundo Fogliatti e Mattos (2007), no regime estacionário de qualquer processo markoviano

$$P_n(t) = P_n \quad \forall n \geq 0.$$

Para um processo representado pelo modelo $M/M/1$ que se encontra nesse regime, substituindo na fórmula (2.17) as taxas λ_n e μ_n por λ e μ, respectivamente, obtém-se

$$P_n = \frac{\lambda^n}{\mu^n} P_0, \quad \forall n \geq 1, \tag{2.26}$$

e de (2.18),

$$P_0 = \left[\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^n \right]^{-1},$$

onde a soma geométrica só converge se $\frac{\lambda}{\mu} < 1$. Neste caso, tem-se:

$$P_0 = 1 - \frac{\lambda}{\mu}. \tag{2.27}$$

O parâmetro ρ definido como

$$\rho = \frac{\lambda}{\mu}$$

é denominado taxa de ocupação/utilização do sistema, que substituindo em (2.26) e (2.27) leva a

$$P_n = \rho^n (1 - \rho), \quad \forall n \geq 0. \tag{2.28}$$

Da Expressão (2.28), observa-se que o número de usuários no sistema segue uma distribuição geométrica modificada de parâmetro $(1-\rho)$ com valor esperado dado por:

$$W = \frac{\rho}{1-\rho}. \tag{2.29}$$

Conforme apresentado na Seção 2.5.4, as medidas de desempenho de um sistema em regime estacionário auxiliam na avaliação da produtividade e no dimensionamento do mesmo. Serão apresentadas a seguir algumas medidas de desempenho correspondentes ao modelo $M/M/1$.

Número médio de usuários no sistema (L)

Seja N a variável aleatória discreta que representa o número de usuários no sistema no regime estacionário, com distribuição de probabilidade $\{P_n\}, n \geq 0$ e valor esperado L. Tem-se, então:

$$L = E[N] = \sum_{n=0}^{\infty} nP_n.$$

Usando a Equação (2.28), segue que

$$L = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1} = (1-\rho)\rho\sum_{n=1}^{\infty}\frac{d\rho^n}{d\rho}.$$

Se $\rho < 1$, então $\sum_{n=0}^{\infty}\rho^n$ converge e então,

$$\sum_{n=1}^{\infty}\frac{d\rho^n}{d\rho} = \frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right).$$

Dessa forma,

$$L = (1-\rho)\rho\frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right) = (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = (1-\rho)\rho\frac{1}{(1-\rho)^2}.$$

De onde se conclui que

$$L = \frac{\rho}{1-\rho},$$

o que confirma a expressão (2.29)

Número médio de usuários na fila (L_q)

Seja N_q a variável aleatória discreta que representa o número de usuários na fila no regime estacionário e L_q seu valor esperado. Então,

$$N_q = \begin{cases} N-1, & \forall N \geq 1, \\ 0, & N = 0, \end{cases}$$

de onde

$$\begin{aligned} L_q = E[N_q] &= \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=1}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n = L - 1 + P_0 \\ &= \frac{\rho}{1-\rho} - 1 + 1 - \rho = \frac{\rho - \rho + \rho^2}{(1-\rho)} = \frac{\rho^2}{(1-\rho)}. \end{aligned}$$

Probabilidade de se ter mais do que k elementos no sistema

Apesar de se tratar de um sistema com capacidade infinita, estas probabilidades são úteis para se avaliar a necessidade de incluir certas comodidades no local reservado para a fila, como cadeiras, banheiros ou outras instalações. Para este sistema particular de fila, vale

$$\begin{aligned} P(N \geq k) &= \sum_{n=k}^{\infty} P_n = \sum_{n=k}^{\infty} \rho^n(1-\rho) = (1-\rho)\sum_{i=0}^{\infty} \rho^{k+1} \\ &= (1-\rho)\rho^k \sum_{i=0}^{\infty} \rho_i = (1-\rho)\rho^k \frac{1}{1-\rho}, \end{aligned}$$

de onde segue que

$$P(N \geq k) = \rho^k.$$

Função de distribuição acumulada do tempo de espera na fila ($W_q(t)$)

Considerando-se o sistema no regime estacionário, seja T_q a variável aleatória contínua que representa o tempo que um usuário qualquer permanece na fila aguardando por atendimento. Esse tempo depende do número de unidades que se encontram à sua frente e do tempo que essas unidades levam para ser atendidas.

Na chegada de um usuário no sistema, podem ser identificados dois eventos mutuamente excludentes:

1. O sistema está vazio, então, $T_q = 0$;

2. Há n elementos no sistema, $n > 0$, então, $T_q > 0$.

Seja $W_q(t)$ a função de distribuição acumulada de T_q que expressa a probabilidade de um usuário qualquer aguardar na fila no máximo um tempo $t \geq 0$. Então,

$$W_q(t) = P(T_q \leq t).$$

$$W_q(0) = P(T_q \leq 0) = P(N = 0) = P_0 = 1 - \rho$$

e para $t > 0$,

$$W_q(t) = \sum_{n=0}^{\infty} P(n \text{ usuários no sistema e os } n \text{ serviços completados até } t).$$

O evento "n serviços completados até t" é equivalente ao evento "o tempo para completar n serviços é menor do que t".

Seja $T_{(n)}$ a variável aleatória contínua que representa a soma dos tempos de atendimento de n usuários consecutivos. Os tempos de serviço são independentes e exponencialmente distribuídos com taxa μ, consequentemente $T_{(n)}$ segue uma distribuição de Erlang de parâmetros n e μ. Então, $\forall t \geq 0$,

$$\begin{aligned}
W_q(t) &= W_q(0) + \sum_{n=1}^{\infty} P[(n \text{ usuários no sistema} \cap (T_{(n)} \leq t))] \\
&= P_0 + \sum_{n=1}^{\infty} P_n P[T_{(n)} \leq t | n \text{ usuários no sistema}] \\
&= (1-\rho) + \sum_{n=1}^{\infty} [\rho^n (1-\rho)] \left(\int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx \right) \\
&= (1-\rho) + \rho(1-\rho) \int_0^t \left(\mu e^{-\mu x} \sum_{n=1}^{\infty} \frac{(\lambda x)^{n-1}}{(n-1)!} \right) dx \\
&= (1-\rho) + \rho(1-\rho)\mu \int_0^t e^{-(\mu-\lambda)x} dx.
\end{aligned}$$

Sabendo que $(\mu - \lambda) = \mu(1-\rho)$, tem-se:

$$\begin{aligned}
W_q(t) &= (1-\rho) + \rho - \rho e^{-(\mu-\lambda)t} \\
&= 1 - \rho e^{-(\mu-\lambda)t}.
\end{aligned} \tag{2.30}$$

Função de densidade do tempo de espera na fila ($w_q(t)$)

Como se sabe, a função de densidade de probabilidade de uma determinada variável aleatória pode ser obtida como a derivada de sua função de distribuição acumulada. Assim, pela Equação (2.30)

$$w_q(t) = \frac{dW_q(t)}{dt} = \frac{d(1 - \rho e^{-(\mu-\lambda)t})}{dt} = \rho(\mu - \lambda)e^{(\mu-\lambda)t}.$$

Função de distribuição acumulada do tempo de permanência no sistema ($W(t)$)

Com raciocínio análogo ao utilizado para obter $W_q(t)$, obtém-se a função de distribuição acumulada do tempo T de permanência no sistema, $W(t)$:

$$W(t) = 1 - e^{-\mu(1-\rho)t} \quad \forall t \geq 0 \tag{2.31}$$

Da expressão (2.31), observa-se que a variável aleatória T, tempo de permanência no sistema, segue uma distribuição exponencial de parâmetro $\mu(1-\rho)$, com valor esperado dado por

$$E[T] = \frac{1}{\mu(1-\rho)} = \frac{1}{(\mu - \lambda)}. \tag{2.32}$$

Tempo médio de espera na fila (W_q)

O valor esperado de T_q, é dado por:

$$\begin{aligned} W_q = E[T_q] &= \int_0^\infty t w_q(t) dt \\ &= \int_0^\infty t\rho(\mu - \lambda)e^{-(\mu-\lambda)t} dt \\ &= \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu - \lambda}. \end{aligned}$$

Tempo médio de permanência no sistema (W)

Esta média pode ser calculada observando-se que o tempo médio que um usuário qualquer permanece no sistema é igual à soma do tempo médio de espera na fila com o tempo médio de atendimento, ou seja,

$$W = W_q + \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)} + \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

o que confirma a expressão (2.32).

Probabilidade do tempo de espera na fila ser maior do que um tempo $t > 0$

Usando a Equação (2.30), segue que

$$P(T_q > t) = 1 - W_q(t) = \rho e^{-(\mu-\lambda)t}.$$

Fórmulas de Little

Conforme Fogliatti e Mattos (2007), um importante resultado geral que independe de propriedades específicas das distribuições dos tempos entre chegadas e de atendimento estabelece que: "o número médio de usuários num sistema é igual ao produto da taxa média de ingresso pelo tempo médio de permanência de um usuário no mesmo". Este resultado, conhecido como fórmula de Little (LITTLE, 1961 apud FOGLIATTI; MATTOS, 2007), é representado analiticamente por:

$$L = E[\Lambda]W \tag{2.33}$$

onde $E[\Lambda]$ é a taxa média de ingressos no sistema.

Outras relações entre as medidas de desempenho do sistema, também conhecidas como fórmulas de Little, válidas em geral são:

$$W = W_q + E[S] \tag{2.34}$$

onde S é o tempo que um usuário qualquer permanece em atendimento,

$$W_q = \frac{L_q}{E[\Lambda]} \tag{2.35}$$

e

$$L_q = L - E[\Lambda]E[S].$$

As provas são indutivas e a utilidade deste conjunto de relações reside no fato de que o conhecimento de uma das medidas de desempenho implica o conhecimento das outras.

Para o modelo $M/M/1$, tem-se $E[\Lambda] = \lambda$ e $E[S] = \frac{1}{\mu}$, portanto:

$$L = \frac{\lambda}{\mu - \lambda},$$

$$W = \frac{\lambda}{\lambda(\mu - \lambda)} + \frac{1}{\mu} = \frac{1}{\mu - \lambda},$$
$$W_q = \frac{\rho^2}{(1-\rho)\lambda} = \frac{\lambda}{\mu(\mu - \lambda)},$$
$$L_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu - \lambda)}.$$

2.6.2 Modelo $M/M/1/k/FIFO$

O modelo $M/M/1/k/FIFO$ será representado como $M/M/1/k$ para simplificar a notação. Tal modelo é caracterizado por:

- tempos entre chegadas sucessivas e os tempos de atendimento seguem distribuições exponenciais de parâmetros λ e μ, respectivamente;
- existe um único posto de atendimento;
- o usuário é atendido conforme sua ordem de chegada;
- a capacidade do sistema é limitada a k usuários no sistema.

Neste modelo, as chegadas e os atendimentos caracterizam um Processo de Nascimento e Morte. Entretanto, a taxa de ingresso ao sistema, $\lambda_n^{'}$, difere da taxa de chegada para $n \geq k$, tendo em vista a existência da limitação na capacidade do sistema (igual a k). Neste caso, as taxas de ingresso e de atendimento são dadas respectivamente por:

$$\lambda_n^{'} = \begin{cases} \lambda, & 0 \leq n < k, \\ 0, & n \geq k \end{cases}$$

e

$$\mu_n = \mu, \quad \forall n \geq 1.$$

Para o estado de regime estacionário do sistema, tem-se:

$$P_n(t) = P_n, \quad 0 \leq n \leq k.$$

Substituindo, em (2.17), $\lambda_n^{'}$ e μ_n pelos seus respectivos valores, tem-se:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0, \quad 0 < n \leq k.$$

De (2.18), fazendo $\rho = \frac{\lambda}{\mu}$, tem-se:

$$P_0 = \frac{1}{\sum_{n=0}^{k} \rho^n}.$$

A soma finita do denominador sempre converge, porém para valores distintos, dependendo de ρ. Dessa forma,

$$P_0 = \begin{cases} \frac{1}{k+1}, & \text{se } \rho = 1, \\ \\ \frac{1-\rho}{1-\rho^{k+1}}, & \text{se } \rho \neq 1, \end{cases} \tag{2.36}$$

de onde, $\forall\, 0 < n \leq k$

$$P_n = \begin{cases} \frac{1}{k+1}, & \text{se } \rho = 1, \\ \\ \frac{(1-\rho)\rho^n}{1-\rho^{k+1}}, & \text{se } \rho \neq 1, \end{cases} \tag{2.37}$$

Conforme apresentado na Seção 2.5.4, as medidas de desempenho de um sistema em regime estacionário auxiliam na avaliação da produtividade e no dimensionamento do mesmo. Serão apresentadas a seguir algumas medidas de desempenho correspondentes ao modelo $M/M/1/k$.

Número médio de usuários no sistema (L)

Por definição

$$L = E[N] = \sum_{n=0}^{k} nP_n.$$

Usando (2.36) e (2.37), e supondo que $\rho = 1$,

$$L = \sum_{n=0}^{k} n\frac{1}{(k+1)} = \frac{1}{(k+1)} \cdot \frac{k(k+1)}{2} = \frac{k}{2}.$$

Se $\rho \neq 1$,

$$\begin{aligned} L &= \frac{1-\rho}{(1-\rho^{k+1})} \sum_{n=0}^{k} n\rho^n = \frac{(1-\rho)\rho}{(1-\rho^{k+1})} \sum_{n=1}^{k} n\rho^{n-1} \\ &= \frac{(1-\rho)\rho}{(1-\rho^{k+1})} \sum_{n=1}^{k} \frac{d\rho^n}{d\rho} = \frac{(1-\rho)\rho}{(1-\rho^{k+1})} \cdot \frac{d}{d\rho} \sum_{n=0}^{k} \rho^n \\ &= \frac{(1-\rho)\rho}{(1-\rho^{k+1})} \cdot \frac{d}{d\rho} \left(\frac{1-\rho^{k+1}}{1-\rho} \right) \\ &= \frac{(1-\rho)\rho}{(1-\rho^{k+1})} \cdot \frac{[-(1-\rho)(k+1)\rho^k + 1 - \rho^{k+1}]}{(1-\rho)^2} \\ &= \frac{\rho[1 + k\rho^{k+1} - \rho^k(k+1)]}{(1-\rho)(1-\rho^{k+1})}. \end{aligned}$$

Em resumo,

$$L = \begin{cases} \frac{k}{2}, & \text{se } \rho = 1 \\ \\ \frac{\rho[1+k\rho^{k+1}-\rho^k(k+1)]}{(1-\rho)(1-\rho^{k+1})}, & \text{se } \rho \neq 1. \end{cases} \tag{2.38}$$

Número médio de usuários na fila (L_q)

Sabendo que

$$L_q = E[N_q]$$

tem-se que

$$L_q = \sum_{n=1}^{k} (n-1)P_n = \sum_{n=0}^{k} nP_n - \sum_{n=1}^{k} P_n = L - 1 + P_0,$$

sendo que L é dado por (2.38).

Tempo médio de permanência no sistema (W)

Para se usar a fórmula de Little (2.33), deve ser feita uma modificação, pois, por existir limitação do espaço reservado para a fila, rejeições acontecem com taxa λP_k cada vez que o sistema atinge o estado k. Dessa forma, a taxa de ingressos, λ', não coincide com a taxa de chegadas, λ.

A taxa de efetivos ingressos λ', é dada por

$$\lambda' = \lambda - \lambda P_k = \lambda(1 - P_k),$$

que substituída em (2.33) leva a

$$W = \frac{L}{\lambda(1 - P_k)}.$$

Tempo médio de espera na fila: W_q

De (2.34),

$$W_q = W - \frac{1}{\lambda}$$

e utilizando (2.35) tem-se a fórmula equivalente

$$W_q = \frac{L_q}{\lambda(1 - P_k)}.$$

Função de distribuição acumulada para o tempo de espera na fila ($W_q(t)$)

A obtenção desta função segue lógica análoga à aplicada no modelo $M/M/1$, com duas diferenças: a série utilizada na derivação é finita e as probabilidades de haver n elementos no sistema devem sofrer uma correção devido à existência de rejeições. Seja N^* a variável que representa o número de usuários que efetivamente ingressam no sistema. Sua distribuição de probabilidade q_n é a distribuição da variável aleatória N, definida no modelo $M/M/1/\infty/FIFO$, truncada à direita em $n = k$, portanto,

$$q_n = \begin{cases} lP_n, & 0 \leq n \leq k-1, \\ 0, & n \geq k, \end{cases}$$

onde

$$l = \frac{1}{1 - P_k}$$

é a constante normalizadora de forma tal que

$$\sum_{n=0}^{k-1} q_n = 1.$$

Dessa forma,

$$q_n = \frac{P_n}{1 - P_k} \quad n \leq k-1.$$

O evento "n serviços completados até t" é equivalente a "o tempo para completar n serviços é menor do que t". Seja $T_{(n)}$ a variável aleatória que representa a soma dos tempos de atendimentos de n usuários consecutivos. Como cada usuário tem um tempo de serviço exponencialmente distribuído com parâmetro igual a μ, $T_{(n)}$ segue uma distribuição de Erlang de parâmetros n e μ. Então,

$$
\begin{aligned}
W_q(t) &= P(T_q \leq t) \\
&= P(\text{todos os usuários no sistema serem atendidos num tempo menor que } t) \\
&= \sum_{n=0}^{k-1} P(n \text{ usuários no sistema e } n \text{ serviços completados até } t) \\
&= W_q(0) + \sum_{n=1}^{k-1} q_n \int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx \\
&= W_q(0) + \sum_{n=1}^{k-1} q_n \left[1 - \int_0^\infty \mu \frac{(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx \right] \\
&= W_q(0) + \sum_{n=0}^{k-2} q_{n+1} \left[1 - \int_0^\infty \mu \frac{(\mu x)^{n}}{n!} e^{-\mu x} dx \right].
\end{aligned}
$$

Como

$$
\begin{aligned}
\int_0^\infty \mu \frac{(\mu x)^n}{n!} e^{-\mu x} dx &= P(\text{tempo para completar } (n+1) \text{ serviços} \geq t) \\
&= P(\text{completar no máximo } n \text{ serviços até } t) \\
&= \sum_{i=0}^{n} \frac{(\mu t)^i}{i!} e^{-\mu t},
\end{aligned}
$$

tem-se:

$$
\begin{aligned}
W_q &= W_q(0) + \sum_{n=0}^{k-2} q_{n+1} \left[1 - \sum_{i=0}^{n} \frac{(\mu t)^i}{i!} e^{-\mu t} \right] \\
&= 1 - \sum_{n=0}^{k-2} q_{n+1} \sum_{i=0}^{n} \frac{(\mu t)^i}{i!} e^{-\mu t},
\end{aligned}
$$

pois $W_q = q_0$.

Probabilidade de se ter pelo menos k elementos no sistema: $(k \leq k)$

$$
P(N \geq k) = \sum_{n=k}^{k} P_n = \begin{cases} \frac{(k+1-k)}{k+1}, & \text{se } \rho = 1, \\ \\ \rho^k \frac{(1-\rho^{k+1-k})}{1-\rho^{k+1}}, & \text{se } \rho \neq 1, \end{cases}
$$

2.6.2.1 Caso particular $M/M/1/1/FIFO$

Neste caso, o sistema não admite fila, ou seja, enquanto o único servidor estiver ocupado, novos usuários são impedidos de entrar no sistema. Então existem apenas dois estados: $n = 0$ e $n = 1$. Tem-se neste caso, para qualquer ρ,

$$P_1 = \rho P_0.$$

Como $P_0 + P_1 = 1$, tem-se

$$P_0 = \frac{1}{(1+\rho)}$$

e

$$P_1 = \frac{\rho}{(1+\rho)}.$$

2.6.3 Modelo $M/M/c/\infty/FIFO$

No modelo $M/M/c/\infty/FIFO$, de acordo Fogliatti e Mattos (2007), os tempos entre chegadas sucessivas seguem distribuições exponenciais de parâmetro λ e há c servidores, cada um dos quais com tempos de atendimento que seguem distribuições exponenciais, de parâmetro μ. Como as chegadas e os atendimentos neste caso caracterizam Processos de Nascimento e Morte, logo as taxas de chegadas e de atendimento respectivamente são dadas por:

$$\lambda_n = \lambda \quad \forall\, n \geq 0 \tag{2.39}$$

e

$$\mu_n = \begin{cases} n\mu, & \text{se } 1 \leq n < c, \\ c\mu, & \text{se } n \geq c. \end{cases} \tag{2.40}$$

Denotando $r = \frac{\lambda}{\mu}$, a taxa de utilização do sistema é dada por:

$$\rho = \frac{r}{c} = \frac{\lambda}{c\mu}.$$

Substituindo (2.39) e (2.40) em (2.17), obtém-se:

$$P_n = \begin{cases} P_0 \frac{r^n}{n!}, & \text{se } 1 \leq n < c, \\ \\ P_0 \frac{r^n}{c^{n-c}c!}, & \text{se } n \geq c \end{cases}$$

o que implica que

$$\begin{aligned}\sum_{n=0}^{\infty} P_n &= \sum_{n=0}^{c-1} P_n + \sum_{n=c}^{\infty} P_n \\ &= P_0 \sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{P_0}{c!} \sum_{n=c}^{\infty} \frac{r^n}{c^{n-c}} \\ &= P_0 \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!} \sum_{i=0}^{\infty} \rho^i \right).\end{aligned}$$

A soma converge se $\rho < 1$. Nesse caso, tem-se:

$$\begin{aligned}\sum_{n=0}^{\infty} P_n &= P_0 \left[\left(\sum_{n=0}^{c-1} \frac{r^n}{n!} \right) + \frac{r^c}{c!} \cdot \frac{1}{1-\rho} \right] \\ &= P_0 \left[\left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{cr^c}{c!(c-r)} \right) \right].\end{aligned}$$

Como $\sum_{n=0}^{\infty} P_n = 1$, então,

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{cr^c}{c!(c-r)} \right)^{-1} \tag{2.41}$$

Serão apresentadas a seguir algumas medidas de desempenho correspondentes ao modelo $M/M/c$.

Número médio de usuários na fila (L_q)

Para este modelo, tem-se que

$$\begin{aligned}L_q = E[N_q] &= \sum_{n=c}^{\infty} (n-c) P_n \\ &= \frac{P_0 r^{c+1}}{c!c} \cdot \sum_{n=c}^{\infty} \frac{(n-c) r^{n-c-1}}{c^{n-c-1}} \\ &= \frac{P_0 r^{c+1}}{c!c} \cdot \frac{\partial}{\partial(\frac{r}{c})} \sum_{n=0}^{\infty} \left(\frac{r}{c}\right)^n \\ &= \frac{P_0 r^{c+1}}{c!c} \cdot \frac{\partial \left(\frac{1}{1-\frac{r}{c}}\right)}{\partial(\frac{r}{c})} \\ &= \frac{P_0 r^{c+1}}{c!c} \cdot \frac{1}{\left(1-\frac{r}{c}\right)^2} = \frac{P_0 c r^{c+1}}{c!(c-r)^2}.\end{aligned}$$

Usando a Fórmula de Little (2.33) e as relações (2.34) e (2.35), obtêm-se as demais medidas de desempenho:

- **Número médio de usuários no sistema:** $L = r + \left[\frac{r^{c+1}c}{c!(c-r)^2}\right] P_0$.
- **Tempo médio de espera na fila:** $W_q = \frac{r^c\mu}{(c-1)!(c\mu-\lambda)^2} P_0$.
- **Tempo médio de permanência no sistema:** $W = \frac{1}{\mu} + \left[\frac{r^c\mu}{(c-1)!(c\mu-\lambda)^2}\right] P_0$.

Função de distribuição acumulada, $W_q(t)$ do tempo de espera na fila

Considere T_q como sendo o tempo de espera na fila de um usuário qualquer. Como há c postos de atendimento, então T_q será zero quando o número de usuários à frente do usuário considerado for menor ou igual a $c-1$, ou seja,

$$P(T_q = 0) = P(N \leq c-1) = \sum_{n=0}^{c-1} P_n = P_0 \sum_{n=0}^{c-1} \frac{r^n}{n!}.$$

Neste caso, definindo $W_q(t)$ como a função de distribuição acumulada de T_q e usando a Equação (2.41), segue que

$$W_q(0) = P(T_q = 0) = 1 - P_0 \frac{cr^c}{c!(c-r)}.$$

De acordo com Fogliatti e Mattos (2007), o tempo T_q que um usuário aguarda na fila é positivo e no máximo t unidades de tempo se esse usuário encontra à sua frente n usuários com $n \geq c$ e os servidores completam pelo menos $(n-c-1)$ serviços até t. Desta forma, calcular a probabilidade de que ***n* serviços sejam completados até o tempo *t*** é o mesmo que calcular a probabilidade de que **o tempo para completar *n* serviços seja menor do que *t***.

Assim, definindo a variável aleatória

$$T_{(n)}: \text{ soma dos tempos de atendimentos de } n \text{ usuários consecutivos,}$$

então, dado que os tempos de serviço são exponenciais com taxa μ, $T_{(n)}$ segue uma distribuição de Erlang de parâmetros n e μ. Portanto

$$W_q(t) = P(T_q \leq t),$$

o que implica

$$
\begin{aligned}
W_q(t) &= W_q(0) + \sum_{n=c}^{\infty} P(n \text{ usuários no sistema e pelo menos } n-c+1 \text{ serviços completados até } t) \\
&= W_q(0) + P_0 \sum_{n=c}^{\infty} \frac{r^n}{c^{n-c}c!} \int_0^{\infty} \frac{\mu c(\mu c x)^{n-c}}{(n-c)!} e^{-\mu c x} dx \\
&= W_q(0) + P_0 \frac{r^c}{(c-1)!} \int_0^{\infty} \mu e^{-\mu c x} \sum_{n=c}^{\infty} \frac{(\mu r x)^{n-c}}{(n-c)!} dx \\
&= W_q(0) + P_0 \frac{r^c}{(c-1)!} \int_0^{t} \mu e^{-\mu(c-r)x} dx \\
&= W_q(0) + P_0 \frac{r^c}{(c-1)!} \cdot \frac{(1 - e^{-\mu(c-r)t})}{(c-r)} = 1 - P_0 \frac{r^c}{c!(1-\rho)} e^{-(c\mu-\lambda)t}
\end{aligned}
$$

2.6.4 Modelo $M/M/c/k/FIFO$

De acordo com a descrição inicial de um modelo de filas, o modelo $M/M/c/k/FIFO$ é caracterizado por:

- os tempos entre chegadas sucessivas seguem distribuições exponenciais de parâmetro λ;
- os tempos de atendimento em cada posto de atendimento seguem distribuições exponenciais, de parâmetro μ;
- há c servidores
- O sistema comporta k clientes;
- A ordem de acesso de usuários ao serviço é a ordem das chegadas dos mesmo ao sistema.

Como nos casos anteriores, trata-se de um Processo de Nascimento e Morte, sendo que a taxa de ingresso ao sistema, $\lambda_n^{'}$ não é igual à taxa de chegada λ para $n \geq k$, pois, como foi dito, o sistema tem capacidade limitada. As taxas de ingresso e de atendimento são dadas, respectivamente, por:

$$
\lambda_n^{'} = \begin{cases} \lambda, & 0 \leq n < k, \\ \\ 0, & n \geq k \end{cases} \tag{2.42}
$$

e

$$\mu_n = \begin{cases} n\mu, & 1 \leq n < c, \\ \\ c\mu, & c \leq n \leq k. \end{cases} \tag{2.43}$$

A seguir será apresentada a caracterização do sistema no regime estacionário. Denotando $r = \frac{\lambda}{\mu}$ e substituindo (2.42) e (2.43) em (2.17), obtém-se:

$$P_n = \begin{cases} \left(\frac{r^n}{n!}\right) P_0, & 1 \leq n \leq c-1, \\ \\ \left(\frac{r^n}{c!c^{n-c}!}\right) P_0, & c \leq n \leq k. \end{cases}$$

Então,

$$\begin{aligned} \sum_{n=0}^{k} P_n &= P_0 \sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{P_0}{c!} \sum_{n=c}^{k} \frac{r^n}{c^{n-c}} \\ P_0 &= \left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!} \sum_{i=0}^{k-c} \left(\frac{r}{c}\right)^i \right]. \end{aligned}$$

Como

$$\sum_{n=0}^{k} P_n = 1,$$

$$P_0 = \begin{cases} \left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c(k-c+1)}{c!} \right]^{-1}, & \text{se } \frac{r}{c} = 1 \\ \\ \left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c \left(1 - [\frac{r}{c}]^{k-c+1}\right)}{c! \left(1 - \frac{r}{c}\right)} \right], & \text{se } \frac{r}{c} \neq 1 \end{cases} \tag{2.44}$$

As medidas de desempenho para este sistema serão mostradas a seguir.

Número médio de usuários na fila (L_q)

Para este sistema de filas, vale

$$L_q = E[N_q] = \sum_{n=c}^{k} (n-c) P_n = P_0 \sum_{n=c}^{k} (n-c) \frac{r^n}{c!c^{n-c}} = \frac{P_0 r^{c+1}}{c!c} \sum_{n=c}^{k} (n-c) \left(\frac{r}{c}\right)^{n-c-1}.$$

De acordo com (2.44), vamos considerar dois casos. Primeiramente, se $\frac{r}{c} = 1$, então,

$$L_q = \frac{P_0 r^c}{c!} \sum_{n=c}^{k} (n-c) = \frac{P_0 r^c}{c!} \sum_{i=0}^{k-c} i = \frac{P_0 r^c}{c!} \cdot \frac{(k-c+1)(k-c)}{2}.$$

Caso contrário, se $\frac{r}{c} \neq 1$, então,

$$\begin{aligned} L_q &= \frac{P_0 r^{c+1}}{c!c} \sum_{n=c}^{k} \frac{d}{d(\frac{r}{c})} \left(\frac{r}{c}\right)^{n-c} \\ &= \frac{P_0 r^{c+1}}{c!c} \cdot \frac{d}{d(\frac{r}{c})} \sum_{i=0}^{k-c} \left(\frac{r}{c}\right)^{i} \\ &= \frac{P_0 r^{c+1}}{c!c} \cdot \frac{d\left(\frac{1-(\frac{r}{c})^{k-c+1}}{1-(\frac{r}{c})}\right)}{d(\frac{r}{c})} \\ &= \frac{P_0 r^{c+1}}{c!c} \cdot \frac{\left([(\frac{r}{c})-1](k-c+1)(\frac{r}{c})^{k-c}+1-(\frac{r}{c})^{k-c+1}\right)}{(1-(\frac{r}{c}))} \end{aligned}$$

Número médio de usuários no sistema (L)

É possível desenvolver a fórmula do número médio de usuários na fila da seguinte forma. Por definição,

$$L_q = \sum_{n=c}^{k} (n-c)P_n = \sum_{n=c}^{k} nP_n - c\sum_{n=c}^{k} P_n.$$

Como

$$\sum_{n=0}^{k} P_n = 1 \quad \text{e} \quad L = \sum_{n=0}^{k} nP_n,$$

segue que

$$\begin{aligned} L_q &= \left(L - \sum_{n=0}^{c-1} nP_n\right) - c\left(1 - \sum_{n=0}^{c-1} P_n\right). \\ &= L - \sum_{n=0}^{c-1} nP_n - c + c\sum_{n=0}^{c-1} P_n = L - \sum_{n=0}^{c-1} (n-c)P_n - c, \end{aligned}$$

de onde se conclui que

$$L = L_q + c + \sum_{n=0}^{c-1} (n-c)P_n.$$

Tempo médio de espera na fila (W_q)

Utilizando as fórmulas (2.33) e (2.35) e lembrando que existe limitação no espaço físico reservado para a fila, tem-se

$$W_q = \frac{L_q}{\lambda'}$$

com $\lambda' = \lambda(1 - P_k)$.

Tempo médio de permanência no sistema (W)

Por fim, vê-se que o tempo médio de permanência no sistema é dado por

$$W = \frac{L}{\lambda^{'}},$$

com $\lambda^{'} = \lambda(1 - P_k)$.

Caso Particular: $M/M/c/c/FIFO$

Este caso considera que a capacidade do sistema é igual ao número de postos de atendimento, ou seja, o cliente chega e só entrará no sistema se algum posto de atendimento estiver livre. Como não é permitida a formação de fila, as taxas de ingresso e de atendimento do sistema são dadas respectivamente por:

$$\lambda_n^{'} = \begin{cases} \lambda, & 0 \leq n < c, \\ \\ 0, & n \geq c, \end{cases}$$

e

$$\mu_n = n\mu, \quad 1 \leq n \leq c.$$

Denotando $r = \frac{\lambda}{\mu}$ e substituindo essas taxas em (2.17), tem-se:

$$P_n = \frac{r^n}{n!}P_0, \quad 0 < n \leq c,$$

de onde,

$$P_0 = \left[\sum_{n=0}^{c} \frac{r^n}{n!}\right]^{-1}$$

e

$$P_n = \frac{\left(\frac{r^n}{n!}\right)}{\sum_{i=0}^{c} \frac{r^i}{i!}}. \quad 0 \leq n \leq c$$

A probabilidade do sistema se encontrar lotado,

$$P_c = \frac{\left(\frac{r^c}{c!}\right)}{\sum_{i=0}^{c} \frac{r^i}{i!}},$$

é conhecida como a fórmula de perda de Erlang (Erlang, 1917) e corresponde à percentagem de usuários rejeitados pela limitação física dos sistema.

3 Metodologia

Este trabalho dividiu-se em duas partes: na primeira parte foi realizado um estudo teórico do conteúdo proposto e, num segundo momento, foi feita uma aplicação da Teoria das Filas num conjunto de dados real.

No que se refere ao estudo teórico, inicialmente, foi preciso fazer uma revisão aprofundada das distribuições de proabilidade a serem utilizadas: Poisson e Exponencial. Em seguida, foi feito um estudo sobre Processos estocásticos, com ênfase nas Cadeias de Markov e Processos de Poisson. Logo após, o foco foi direcionado em cima dos principais conceitos e dos modelos básicos de filas Markovianas. Toda essa revisão teórica foi feita tomando como base os livros Ross (2007) e Fogliatti e Mattos (2007).

A segunda parte é referente a aplicação da Teoria de Filas e cálculo das medidas de desempenho. Várias aplicações foram cogitadas, mas descartadas pela dificuldade de se colher os dados. Por fim, decidiu-se modelar os atendimentos em uma casa lotérica localizada na cidade de Cubati-PB. O estabelecimento utilizado foi uma casa lotérica constituída por dois postos de atendimento, a qual se encaixa no modelo de fila $M/M/2/\infty/FIFO$. O estabelecimento era motivo de constantes reclamações por parte de seus clientes, devido à demora que os mesmos levavam para serem atendidos em dias de pagamento do benefício Bolsa Família do Governo Federal. Com a definição do problema, foram escolhidos dois dias no mês, considerados pela gerência do estabelecimento representativos do comportamento do sistema, para a coleta dos dados. O primeiro dia é referente ao funcionamento do sistema em situação normal, já no segundo dia, havia pagamento do Bolsa Família, o que logicamente causa aumento na fila de espera pelo serviço. Foram observados os tempos entre chegadas sucessivas dos clientes e os tempos que cada um levava em atendimento. Após o colhimento dos dados, foram calculadas as taxas de chegadas sucessivas e de atendimento referentes aos dois dias e suas respectivas medidas de desempenho, no intuito de compará-las para se obter conclusões sobre o funcionamento do sistema e consequentemente propor melhorias ao mesmo. Vale salientar que todos os cálculos e gráficos presentes neste trabalho foram feitos utilizando o *software* R, que pode ser obtido

gratuitamente em http://www.r-project.org/.

Uma vez que os dados foram coletados, foi feito o teste Qui-Quadrado de aderência para saber se era plausível o uso de filas markovianas, ou seja, se o número de chegadas e os atendimentos eram satisfatoriamente modelados por uma distribuição de Poisson. Isso é equivalente a se ter tempos entre chegadas e de atendimentos exponenciais.

É importante que se comente sobre o quanto o estudo da Teoria de Filas é complexo e sobre a dificuldade encontrada na coleta dos dados, já que às vezes, pessoas ou entidades não têm interesse que um estudo dessa natureza seja feito, por receio de resultados insatisfatórios.

4 Resultados e Discussões

Neste trabalho, propusemos aplicar a teoria estudada a um conjunto de dados real. Inicialmente, foi cogitado usar algum sistema de atendimento do SAMU ou de algum posto de saúde. No entanto, tais aplicações são muito complicadas, no que se refere à coleta dos dados. Além de precisar da aprovação do comitê de ética, precisaríamos que os órgãos competentes aceitassem que tal estudo fosse feito. A seguir, serão apresentados os resultados do estudo desenvolvido.

Os dados tratam-se de tempos entre chegadas e de atendimento. Tendo em vista que o tempo é uma variável contínua, torna-se impossível coletar os dados com 100% de precisão, mas a diferença entre os dados coletados e os dados exatos não comprometem o estudo, já que esta diferença é de décimos de segundo e numa análise estatística sempre há um fator aleatório (erro) que deve comportar tais desvios.

Como já dito anteriormente, observou-se durante dois dias o fluxo de pessoas em uma lotérica na cidade de Cubati-PB e anotou-se o tempo entre as chegadas dessas pessoas ao estabelecimento e quanto tempo elas passaram em atendimento no sistema. Foi considerado um dia em que havia pagamento do Bolsa Família e outro dia que não havia o pagamento.

4.1 Modelagem do Sistema em Situação Normal de Funcionamento

Vale lembrar que a casa lotérica considerada tinha dois caixas de atendimento, o que nos leva a considerar o modelo com dois postos de serviço. Quanto à limitação do sistema, qualquer cliente que chegue antes do horário de fechamento poderia ficar esperando para ser atendido e além disso, a casa lotérica, tendo uma "fila única" atende primeiro quem chega primeiro (FIFO). Assim, a ideia é considerar um modelo $M/M/2$, mas para tanto, precisamos ver a adequabilidade de se considerar que as chegadas e os

atendimentos ocorrem segundo um processo de Poisson. Para verificar se tal suposição era plausível, propôs-se o uso de um teste Qui-quadrado de aderência, cuja descrição pode ser encontrada em livros de estatística, tais como Bussab e Morettin (2002).

Primeiramente as hipóteses formuladas para a execução deste teste foram:

$$\begin{cases} H_0: & \text{As chegadas do clientes se adequam a uma distribuição de Poisson} \\ H_1: & \text{As chegadas do clientes não se adequam a uma distribuição de Poisson} \end{cases}$$

As frequências do número de chegadas por minuto foram calculadas e apresentadas na Tabela 1, no intuito de calcular-se a estatística $X^2 = \frac{(O_i - O_e)^2}{O_e}$ e compará-la com o quantil de uma $\chi^2_{(2;95\%)}$.

Tabela 1: Frequências observadas e esperadas do número de chegadas por minuto e cálculo da estatística X^2 para um dia normal de funcionamento

Nº de chegadas por minuto	Frequência observada O_i	Frequência esperada O_e	$\frac{(O_i - O_e)^2}{O_e}$
0	305	307,67	0,02317
1	76	69,66	0,57703
2	6	7,74	0,39116
Total	387	385,07	0,99136

Tivemos que $\chi^2 = 0,99136 < \chi^2_{(1;95\%)} = 5,991$, logo, aceitamos a hipótese nula, portanto, não há evidências significativas para não se levar em consideração que os o número de chegadas por minuto segue uma distribuição de Poisson ao nível de 5% de significância.

Uma vez que foi estabelecido o modelo de fila que será utilizado, serão estimados os parâmetros do modelo (λ e μ) e calculadas as medidas de desempenho para o mesmo. As medidas de desempenho para dias normais de funcionamento foram calculadas e são mostradas na Tabela 2. Podemos ver, entre outras coisas, que a taxa de chegadas é menor que a taxa de atendimento e também que a probabilidade de formação de fila é baixa. Outro ponto que vale a pena ser destacado é que $\rho < 1$, o que é uma condição a considerada no desenvolvimento da teoria apresentada.

Tabela 2: Estimativa dos parâmetros e medidas de desempenho para dias normais de funcionamento

Medidas de desempenho	Valores
Taxa de chegadas (λ)	0,23 clientes/min
μ_1	0,55 clientes/min
μ_2	0,45 clientes/min
Taxa de atendimento (μ)	0,5 clientes/min
$c\mu$	1 cliente/min
ρ	0,23
P_0	0,63
L_q	0,03 clientes
L	0,49 clientes
W_q	0,12 min
W	2,12 min
Probabilidade de haver fila $P(N \geq 2)$	0,08 (8%)

Tendo em vista a observação das medidas de desempenho apresentadas na Tabela 2, foi feita uma simulação do sistema com apenas um posto de atendimento para o caso de dias normais de funcionamento. As medidas de desempenho para este caso são apresentadas na Tabela 3. Comparando com os resultados apresentados anteriormente na Tabela 2, vemos que há um aumento considerável da probabilidade de se formar uma fila. Além disso, o número esperado de usuários no sistema quase duplicou e o tempo médio de espera aumentou em torno de 73%.

Tabela 3: Estimativas dos parâmetros e medidas de desempenho para um dia normal de funcionamento, considerando a existência de apenas um posto de serviço

Medidas de desempenho	Valores
Taxa de chegadas (λ)	0,23 clientes/min
Taxa de atendimento(μ)	0,5 clientes/min
ρ	0,46
P_0	0,54
L_q	0,39 clientes
L	0,85 clientes
W_q	1,7 min
W	3,7 min
Probabilidade de haver fila $P(N \geq 1)$	0,46=46%

4.2 Modelagem do Sistema em Dias de Pagamento do Bolsa Família

Agora as análises feitas na seção anterior serão refeitas para o caso em que os dados foram colhidos em dia de pagamento do benefício concedido pelo Governo Federal a algumas famílias de baixa renda, o Bolsa Família. Também aqui será considerado um modelo $M/M/2$, pelas mesmas razões consideradas anteriormente.

Como feito na seção anterior as hipóteses formuladas para a execução deste teste foram:

$$\begin{cases} H_0: & \text{As chegadas do clientes se adequam a uma distribuição de Poisson} \\ H_1: & \text{As chegadas do clientes não se adequam a uma distribuição de Poisson} \end{cases}$$

As frequências do número de chegadas por minuto foram calculadas e apresentadas na Tabela 4, no intuito de calcular-se a estatística $X^2 = \frac{(O_i - O_e)^2}{O_e}$ e compará-la com o quantil de uma $\chi^2_{(3;95\%)}$.

Tabela 4: Frequências observadas e esperadas do número de chegadas por minuto e cálculo da estatística X^2

Nº de chegadas por (min)	Freq. observada O_i	O_i	Freq. esperada O_e	$\frac{(O_i-O_e)^2}{O_e}$
0	88	88	80,58	0,00417
1	46	46	53,72	1,10943
2	15	15	17,38	0,32591
3	6	9	4,74	3,82861
4	3			
Total	158	158	156,42	5,26812

Tivemos que $\chi^2 = 5,26812 < \chi^2_{(2;95\%)} = 7,815$, logo aceitamos a hipótese nula, portanto não há evidências significativas para não se levar em consideração que o número de chegadas por minuto segue uma distribuição de Poisson ao nível de 5% de significância.

Após aceitarmos que o modelo $M/M/2$ é adequado para modelar os dados, devemos estimar seus parâmetros (λ e μ) e calcular as medidas de desempenho associadas ao mesmo. Tais medidas, calculadas para dias de funcionamento do estabelecimento com pagamento do Bolsa Família, foram calculadas e são mostradas na Tabela 5.

Tabela 5: Medidas de desempenho para dias de pagamento do Bolsa Família

Medidas de desempenho	Valores
Taxa de chegada de clientes (λ)	0,67 clientes/min
μ_1	0,67 clientes/min
μ_2	0,57 clientes/min
Taxa média de atendimento em cada posto (μ)	0,62 clientes/min
$c\mu$	1,24 clientes/min
ρ	0,54
P_0	0,30
L_q	0,45 clientes
L	1,53 clientes
W_q	0,68 min
W	2,29 min
Probabilidade de haver fila $P(N \geq 2)$	0,624

Podemos ver, pelos valores apresentados na Tabela 5 que, como era de se esperar, o número médio de clientes no sistema aumentou em relação aos dias em que não há pagamento do benefício. Obviamente, também aumentou o tempo médio de permanência no sistema. Podemos também observar que a razão entre a taxa de chegadas e de atendimento (ρ) permanece abaixo de 1, o que é uma condição fundamental para o bom funcionamento do sistema.

Devido à grande frequência de assaltos contra agências bancárias, correspondentes bancários, agências dos Correios e casas lotéricas nas cidades do interior do estado, a casa lotérica considerada no estudo contava com uma quantidade relativamente reduzida de dinheiro. Assim, na grande maioria das vezes, faltava dinheiro em um dos caixas de atendimento do estabelecimento em dias de pagamento do Bolsa Família, ou seja, na verdade, o sistema operava basicamente com apenas um posto de serviço, neste caso, portanto, foi feita uma simulação do sistema com apenas um posto de atendimento. Foram calculadas as taxas de chegadas (λ) e de atendimentos (μ) e neste caso observou-se que $\rho = \frac{\lambda}{\mu} > 1$. Tal fato torna impossível o cálculo das medidas de desempenho, pois a soma geométrica usada no cálculo dessas medidas não converge. O que acontece neste caso, é que o sistema sofre uma super lotação, pois chegam mais usuários do que o sistema pode atender, daí um dos motivos das constantes reclamações referentes ao estabelecimento.

4.3 Resumo Geral

Veja o resumo geral e um comparativo das medidas de desempenho na situação normal de funcionamento e em dias de pagamento do Bolsa Família respectivamente na Tabela 6.

Podemos perceber um aumento bastante significativo na medidas de desempenho quando se trata de dias de pagamento do Bolsa Família, no entanto, esse aumento não torna o sistema fora de controle, com a observação de que o mesmo está em perfeito funcionamento, ou seja, com os dois caixas de serviço funcionando em paralelo.

Tabela 6: Comparação entre as Medidas de Desempenho

M.D	S.P.B.F	C.P.B.F	% de Aumento)
λ	0,23 clientes/min	0,67 clientes/min	191,30%
μ	0,50 clientes/min	0,62 clientes/min	24%
$c\mu$	1,00 clientes/min	1,24 clientes/min	24%
ρ	0,23	0,54	129,17%
P_0	0,63=63%	0,30=30%	-33%
L_q	0,03 clientes	0,45 clientes	1400%
L	0,49 clientes	1,53 clientes	212,24%
W_q	0,12 min	0,68 min	466,67%
W	2,12 min	2,29 min	8,02%
$P(N \geq 2)$	0,08=8%	0,624=62,4%	54,40%

5 Conclusões

No que se refere ao estudo teórico da Teoria das Filas, pode-se dizer que foi bastante importante, visto que foi possível solidificar alguns conceitos vistos superficialmente na graduação, bem como aprender coisas novas. O desenvolvimento deste estudo requeria um bom conhecimento de cálculo (derivadas, integrais e séries), o que resultou num crescimento da maturidade matemática, que é fundamental para aqueles que queiram fazer uma pós-graduação na área de Estatística.

Quanto à aplicação, foi muito gratificante poder aplicar a teoria vista em um problema real, o que justifica horas a fio de estudos. A coleta dos dados teve um nível de dificuldade maior do que o esperado, pois não havia muito tempo disponível. Além disso, como já foi descrito, alguns dos lugares onde buscamos realizar a aplicação não se mostraram disponíveis. Muitos lugares, não gostam de divulgar que o seu sistema de filas não é ideal, no sentido de que tenha a taxa de serviço muito pequena em relação à taxa de chegadas.

O que os dados considerados apontaram, como já era de se esperar, é que há uma diferença considerável no comportamento do sistema na situação normal de funcionamento e na situação onde há pagamento do Bolsa Família. Isto é reforçado pelo estudo, pois, temos um aumento no tempo médio que o usuário fica no sistema, no número médio de usuários no sistema e na probabilidade de formação de fila.

Apesar da diferença apresentada entre as medidas de desempenho nas duas situações, pode-se concluir que o sistema se comporta muito bem em ambas desde que ele opere efetivamente com os dois caixas, pois, se levarmos em consideração que o estabelecimento só opera com apenas um posto de serviço na maioria das vezes em dias de pagamento do Bolsa Família, que é o que realmente acontece, o sistema fica super lotado e se torna impossível controlar o congestionamento de clientes no estabelecimento, neste caso, portanto, os clientes têm razão em reclamar a respeito da espera pelo atendimento.

Em dias normais de atendimento, tendo em vista que o número médio de usuários na fila é 0,03 clientes, logo se torna dispensável um posto de atendimento para este caso, já

que na simulação feita com o funcionamento de um só caixa de serviço, o sistema também se comporta muito bem, levando em consideração o tempo médio de permanência no sistema para este caso que é de 3,7min, sendo um tempo tolerável de espera. Já em dias de pagamento do Bolsa Família o ideal era que o sistema operasse com sua capacidade máxima, que seria o funcionamento dos dois postos de atendimento, logo, para que isso ocorra, foi sugerido a gerência do sistema a correção do problema da falta de dinheiro nos caixas de serviço, com isso, o sistema atenderia com qualidade as necessidades de seus usuários além de gerar lucros maiores para o estabelecimento, pois quanto mais clientes atendidos, mais movimentações financeiras são feitas.

De maneira geral, o problema da casa lotérica não está no atendimento, e sim em questões internas como: falta de dinheiro nos caixas de serviço e a rede de computadores fora de conexão. Com a correção desses problemas é possível controlar o fluxo de usuários que entram e saem do estabelecimento, evitando assim perca de clientes e declínio nos lucros da entidade.

Pôde-se concluir também neste trabalho que a Teoria de filas é uma importante ferramenta tratando-se de controle de fluxos de usuários em sistemas com fila, já que através deste estudo pode-se observar detalhadamente o comportamento desses sistemas e ainda propôr melhorias para um bom funcionamento do mesmo.

Referências

ADAMS, W. F. Road traffic considered as a random series. *J. Inst. Civil Eng.*, v. 4, p. 121–130, 1936.

BAILEY, N. T. J. On queueing process with bulk service. *Journal of Royal Statistical Society, Serie B*, v. 16, 1954.

BITRAN, G.; MORABITO, R. Open queueing networks: Optimization and performace evaluation models for discrete manufacturing systems. *Production and Operations Management*, v. 52, p. 163–193, 1996.

BUSSAB, W.; MORETTIN, P. *Estatística Básica*. 5a. ed. [S.l.]: Editora Atual, 2002.

COBHAM, A. Priority assignment in waiting-line problems. *Journal of the Operations Research Society of America*, 1954.

COURTOIS, P. J. *Decomposability: Queueing and Computer System Applications*. [S.l.]: New York: Academic Press, 1977.

DAIGLE, J. *Queueing Theory for Computer Communications*. [S.l.]: New York: Addison-Wesley, 1991.

EVERETT, J. L. Seaport operation as a sthochastic process. *The Journal of the Operations Research Society of America*, n. 1, 1953.

FOGLIATTI, M. C.; MATTOS, N. M. C. *Teoria das Filas*. [S.l.]: Editora Interciência, 2007.

FONTANELLA, G. C.; MORABITO, R. Analyzing the tradeoff between investing in service channels and satistying the targeted user service for brazilian internet providers. *International Transactions in Operational Research*, v. 9, n. 3, p. 247–259, 2001.

GELEMBE, E.; PUJOLLE, G. *Introduction to Queueing Networks*. [S.l.]: Paris: Wiley, 1987.

HILLIER, F. S.; LIEBERMAN, G. J. *Introduction to Operations Research*. 12ª. ed. [S.l.]: Holden-Day Inc., 1974.

KENDALL, D. G. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chains. *Ann. Math. Statist.*, v. 24, p. 338–354, 1953.

LITTLE, J. D. C. A proof for the queueing formula $L = \lambda W$. *Operations Research*, v. 9, p. 383–387, 1961.

MAGALHÃES, M. N. *Introdução à Rede de Filas*. [S.l.]: ABE Associação Brasileira de Estatística, 1996.

MORSE, P. M. *Queues, Inventories and Maintenance*. [S.l.]: New York: John-Wiley & Sons, 1962.

NOVAES, A. G. N. *Pesquisa Operacional e Transportes: Modelos Probabilísticos*. [S.l.]: Mc Graw Hill do Brasil Ltda, 1975.

PRABHU, N. U. *Queues and Inventories*. [S.l.]: New York: Wiley, 1965.

RAMAMOORTHY, C. V. Discrete markov analysis of computer programs. *Proc. ACM National Conference*, p. 386–392, 1965.

ROSS, S. M. *Introduction to Probability models*. 9a. ed. [S.l.]: Academic Press, 2007.

TANNER, J. C. The delay to pedestrians crossing a road. *Biometrika*, v. 38, p. 383–392, 1951.

WALRAND, J. *An Introduction to Queueing Networks*. [S.l.]: New Jersey: Prentice Hall, Englewood Cliffs, 1988.