Emily Hattori

Simulação de Sistema de Filas

Niterói - RJ, Brasil

20 de julho de 2023

Universidade Federal Fluminense

Emily Hattori

Simulação de Sistema de Filas

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Profa. Dra. Jessica Kubrusly

Niterói - RJ, Brasil

20 de julho de 2023

Universidade Federal Fluminense

Emily Hattori

Simulação de Sistema de Filas

Monografia de Projeto Final I de Graduação sob o título *"Simulação de Sistema de Filas"*, defendida por Emily Hattori e aprovada em 20 de julho de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Jessica Kubrusly
Departamento de Estatística – UFF

Profa. Dra. Karina Yuriko Yaginuma
Departamento de Estatística – UFF

Prof. Dr. Antonio Augusto de Aragão Rocha
Departamento de Ciência da Computação – UFF

Niterói, 20 de julho de 2023

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

H366s Hattori, Emily
Simulação de Sistema de Filas / Emily Hattori. - 2023.
57 f.: il.

Orientador: Jessica Quintanilha Kubrusly.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Sistemas de filas. 2. Simulação de números pseudoaleatórios. 3. Processo de Nascimento e Morte. 4. M/M/1. 5. Produção intelectual. I. Kubrusly, Jessica Quintanilha, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O trabalho propõe comparar resultados teóricos de sistemas de filas com resultados simulados. Para isso foram estudadas diversas medidas de desempenho para diferentes sistemas de filas, e essas medidas foram estimadas também a partir da simulação. Os resultados mostraram a convergência das estimativas pela simulação para os valores teóricos, indicando um bom desempenho do algoritmo de simulação. Além disso, foi possível sugerir medidas de desempenho não conhecidas pela teoria, como por exemplo o percentual médio de clientes perdidos com tamanho de fila limitado.

Palavras-chaves: Sistemas de filas; Simulação de números pseudoaleatórios; Processo de Nascimento e Morte; M/M/1/∞/*FIFO*; M/M/1/k/*FIFO*.

Dedicatória

Dedico este trabalho a Deus, que sustenta toda a minha vida.

Agradecimentos

Agradeço, primeiramente, à minha orientadora Jessica por ter me guiado com muita excelência. O tema foi sua ideia, que sugeriu em uma de suas aulas que participei em Geração de Números Aleatórios. E sem ela, o trabalho não teria progredido tanto. Agradeço também a Universidade Federal Fluminense por me proporcionar um bom ensino.

Agradeço à família Kanno (Shinji, Raquel, Airy, Aimy e Emily). Eles me abrigaram na cidade de Niterói e me adotaram como filha e irmã. Sem seu acolhimento e amor, não consigo imaginar como teria sido.

Agradeço aos meus pais, Geraldo e Clara, por sempre me apoiar e me ensinar a buscar sempre a verdade e dar o melhor de mim em tudo que eu faço. Agradeço também ao Eliseu por ter me acompanhado e me incentivado, me ouvindo falar de filas por 1 ano.

Agradeço aos meus amigos por sempre orarem por mim. E, por fim, agradeço a Deus por todas as coisas.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Quando um serviço está ocupado e ainda há clientes a ser atendidos, as filas se mostram um importante assunto. Seja de atendimentos em hospitais, de pagamentos no supermercado ou de serviços ao consumidor via telefone, estas estão presentes e comumente denotadas negativamente, já que resultam necessariamente em espera e eventualmente em imprevistos e prejuízos tanto para os clientes quanto para os prestadores de serviços. Isto posto, a motivação deste trabalho é aumentar o aproveitamento dos serviços minimizando o tempo de espera nas filas e adequando aos diferentes contextos, como preferencialidade. E, se possível, melhorar o período do cliente na fila com informações mais precisas sobre o tempo de espera.

O artigo de Yakimov et al. (2017) teve como objetivo descrever alguns programas de modelagem estrutural (ilustração do sistema) e de simulação. A metodologia consistiu em testar os e comparar parâmetros da simulação e dos valores teóricos. Notou-se que houve uma diferença média de 5% entre os diferentes modelos e a modelagem analítica, o que fez o pesquisador concluir que todos os programas tiveram resultados satisfatórios.

A pesquisa de Wang (2019) visa melhorar o critério de prioridades da fila dos pacientes que precisam de transplantes de rins dos Estados Unidos. Os problemas encontrados incluem o abandono das filas, o custo de espera e o sucesso do procedimento. O diferencial deste sistema é a priorização de pacientes com maior risco de morte. Concluiu-se que a maior parte dos pacientes morrem enquanto esperam na lista e que diferentes políticas não afetam a taxa de abandono significantemente.

O estudo de Smith e Nelson (2015) visou analisar o tempo virtual (tempo estimado) de espera de clientes de um sistema que não é estável e que normalmente não corresponde ao tempo verdadeiro de espera, levando muitos clientes ao engano. A simulação e análise das situações permitiu criar estimativas melhores para a métrica de interesse. Percebeu-se que apesar de não haver nada errado com as simulações, há um problema na interpretação e consequentemente na estimativa do tempo virtual de espera. A melhor forma de fazer

isso é utilizando seções do tempo e verificar a variância desse dado, visto que não é um modelo estacionário.

O objetivo geral deste trabalho é estudar sistemas de filas e simulá-los a fim de aplicar em problemas reais. Os objetivos específicos são: estudar algumas soluções teóricas, apresentar as soluções via simulação, comparar as soluções e simular modelos de atendimento pouco estudados sem valores teóricos.

Este trabalho está organizado da seguinte maneira. No Capítulo 2 são apresentadas as metodologias utilizadas no trabalho, como a fundamentação teórica, os diferentes modelos de filas e os métodos de simulação. No Capítulo 3 são apresentados os resultados das simulações e comparações com seus valores teóricos. E, finalmente, as conclusões no Capítulo 4.

2 Materiais e Métodos

A fim de compreender a Teoria de Filas, é necessário entender alguns conceitos. Portanto, neste capítulo, serão apresentados primeiramente os Processos Estocásticos e a cada seção esses processos serão explorados a fundo para ter o fundamento necessário. Assim, a Teoria de Filas poderá ser apresentada e, logo em seguida, como as simulações serão feitas.

2.1 Processos Estocásticos

A partir dos Processos Estocásticos, é possível construir conceitos importantes para a Teoria de Filas, são eles: as Cadeias de Markov, Processo de Nascimento e Morte e Processos de Poisson. As definições estão em Mendonça (2014).

Definição 2.1 *Seja $\mathcal{T}$ um conjunto arbitrário. Um processo estocástico é uma família $\mathcal{Z} = \{\mathcal{Z}_{(t)}, t \in \mathcal{T}\}$, tal que, para cada $t \in \mathcal{T}$, $\mathcal{Z}_{(t)}$ é uma variável aleatória.*

Um processo estocástico é uma sequência de variáveis aleatórias, onde t usualmente representa o tempo. Neste caso, o conjunto dos índices $\mathcal{T}$ é identificado com $\mathbb{N}$ ou $\mathbb{R}^+$. E é geralmente utilizada com o propósito de representar a evolução aleatória de um fenômeno.

Definição 2.2 *O estado dos espaços de um processo estocástico é definido como o conjunto de todos valores possíveis que as variáveis aleatórias $\mathcal{Z}_{(t)}$ podem assumir.*

De acordo com Mendonça (2014), Se $\{W_{(t)},\ t \in \mathcal{T}\}$ é um processo estocástico com espaço de estados S e conjunto de índices $\mathcal{T}$, então:

- Se S for enumerável, o processo é dito discreto ou a valores inteiros. Se S for um intervalo da reta, então dizemos que é um processo a valores reais.
- Se o conjunto de índices $\mathcal{T}$ for enumerável, então dizemos que o processo é a tempo discreto e, em geral, consideramos $\mathcal{T} = \{0, 1, 2, \dots\}$ e usamos $\{W_{(n)}, n \geq 0\}$ em lugar de $\{W_{(t)},\ t \in \mathcal{T}\}$. Se $\mathcal{T} = [0, \infty)$, $W_{(t)}$ é dito um processo a tempo contínuo.

2.1.1 Cadeias de Markov

Um caso particular de processos estocásticos é a Cadeia de Markov. Uma Cadeia de Markov é um processo estocástico que satisfaz:

$$P(W_{(t+1)} = x \mid W_{(1)}, W_{(2)}, ..., W_{(t)}) = P(W_{(t+1)} = x \mid W_{(t)}).$$

Isto é, uma Cadeia de Markov é um processo em que somente a última informação é relevante para a continuidade do processo, enquanto o passado remoto não.

Além disso, é dita como homogênea quando

$$P(W_{(t+1)} = y \mid W_{(t)} = x) = P(W_{(1)} = x \mid W_{(0)} = y)$$

para qualquer $t \geq 0$. Para facilitar a notação, usa-se $P(W_{(t+1)} = y \mid W_{(t)} = x) = P(x, y)$ como a probabilidade de ir do estado x ao y em um passo. Assim, a probabilidade de um estado mudar de x para y não depende do instante t.

Quando temos uma Cadeia de Markov Homogênea no tempo, suas probabilidades de transição entre dois estados quaisquer podem ser representadas matricialmente na forma

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

onde n é o número de estados do processo.

Definição 2.3 *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 $x \leftrightarrow y$.*

Note que 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 2.4 *Uma Cadeia de Markov é dita ser irredutível se todos os estados se comunicam entre si.*

Um exemplo de Cadeia de Markov irredutível é um site no qual pode-se chegar a qualquer outra página a partir de todas as páginas. Ou seja, todas as páginas deste site se comunicam entre si.

2.1.2 Processo de Nascimento e Morte

Um Processo de Nascimento e Morte é mais um caso particular de um processo estocástico. O processo deve ser homogêneo, irredutível e de tempo contínuo. É importante a compreensão deste, dado que os nascimentos e as mortes descrevem muito bem as entradas e saídas de usuários em um sistema de filas.

De acordo com Ross (2014b), os Processos de Nascimento e Morte são Cadeias de Markov com estados $\{0, 1, ...\}$ em que as transições de um estado q apenas pode ir para um estado $q-1$ ou $q+1$.

Suponha que o espaço de estados seja $S = \{0, 1, 2, ..., q, ...\}$. Se houver uma transição do estado q, isso só poderá ocorrer para seus estados vizinhos $(q-1)$ ou $(q+1)$.

- **Nascimento** é o aumento de um estado q para o estado $(q+1)$
- **Morte** é a diminuição do um estado q para o estado $(q-1)$
- O processo de nascimento e morte descreve probabilisticamente como $W_{(t)}$ se desloca à medida que t aumenta.

Os valores de μ_n e λ_n são respectivamente a taxa média de nascimento e de morte no instante t dentro de um intervalo de tempo de tamanho unitário, que descrevem o quanto os estados aumentam ou diminuem dentro de um intervalo de tempo. Estão definidos como:

$$P(q, q+1) = \lambda_q \tag{2.1}$$

$$P(q, q-1) = 1 - \lambda_q = \mu_n \tag{2.2}$$

$$P(q, k) = 0, \text{ para qualquer } k \neq q-1, q+1 \tag{2.3}$$

Assim também pode-se denotar as probabilidades referentes a essas ocorrências de transição para seus vizinhos imediatos:

(i) A probabilidade $P(W_{(s+t)} = n+1 \mid W_{(s)} = n) = \lambda_n t$ do processo mudar do estado n para $n+1$ em t unidade(s) de tempo é definida como a **taxa de chegada**.

(ii) A probabilidade $P(W_{(s+t)} = n-1 \mid W_{(s)} = n) = \mu_n t$ do processo mudar do estado n para $n-1$ em t unidade(s) de tempo é definida como a **taxa de saída**,

2.1.3 Processo de Poisson

O Processo de Poisson é mais um caso particular dos processos estocásticos. Ademais, é um caso particular do processo de nascimento e morte, da seção anterior.

Definição 2.5 *De acordo com Ross (2014b), um processo estocástico W é chamado de processo de contagem se:*

(i) $W_{(t)} \geq 0 \ \forall t$.

(ii) $W_{(t)}$ *é uma variável aleatória assumidos valores em* $\mathbb{N}$.

(iii) Se $s < t$, *então* $W_{(s)} \leq W_{(t)}$.

(iv) Para $s < t$, $W_{(t)} - W_{(s)}$ *é o incremento do processo no intervalo* $(s, t]$.

Note que um Processo de Contagem é um caso particular de um Processo de Nascimento e Morte com taxa de mortalidade $\mu_n = 0$.

Definição 2.6 *O processo de contagem* $\{W_{(t)},\ t \geq 0\}$ *é um Processo de Poisson com parâmetro* λ, $\lambda > 0$ *se:*

(i) $W_{(0)} = 0$ *(o estado inicial do sistema é o estado 0).*

(ii) O processo tem incrementos independentes, isto é, $W_{(t)} - W_{(s)}$ *é independente de* $W_{(u)} - W_{(t)}$ *quais quer que sejam* $s < t < u$.

(iii) Os incrementos em qualquer intervalo de comprimento t *tem distribuição Poisson com média* λt. *Isto é, para todo* s, $t \geq 0$

$$P(W_{(t+s)} - W_{(s)} = x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}.$$

Note que da condição (iii) segue que o processo Poisson tem incrementos estacionários, ou seja, os incrementos não dependem do intervalo de tempo, além disso

$$E[W_{(t)}] = \lambda t.$$

O que explica o motivo de λ ser chamado de taxa do processo.

Um resultado importante também é que o tempo entre ocorrências sucessivas em um Processo Poisson de taxa λ também pode ser vista como uma variável aleatória que é distribuída exponencialmente. A demonstração desse resultado pode ser encontrada na Seção 2.4 do livro "Teoria de Filas Markovianas e Aplicações"de Mendonça (2014).

Seja W um Processo de Poisson com taxa λ. Seja T_1 a variável aleatória definida pelo tempo que o processo demora para sair do estado 0 para o estado 1, então $T_1 \backsim Exp(\lambda)$. De forma geral, se T é o tempo que o processo demora para sair do estado x e ir para o estado $x+1$, então $T \backsim Exp(\lambda)$.

2.2 Teoria de Filas

Vejamos agora a teoria apresentada aplicada aos Sistemas de Filas.

2.2.1 Processo estocástico N

Defina um processo estocástico $N = \{N_{(i)},\ i \in \mathcal{T}\}$, tal que, para cada $i \in \mathcal{T}$, $N_{(i)}$ é uma variável aleatória que representa a quantidade de pessoas presentes em um sistema de filas no tempo i. Dessa forma, o espaço de estados é definido como $S = \{0, 1, 2, ...\}$, que são valores inteiros que $N_{(i)}$ poderá assumir.

Por exemplo, se $N_{(2)} = 3$, há 3 pessoas no sistema no tempo $i = 2$. Esse será o processo estocástico de interesse desta seção.

$N_{(i)}$ é uma Cadeia de Markov visto que é um processo estocástico em que somente a última informação é relevante para a continuidade do processo, além disso, suponha que satisfaz as seguintes suposições:

- **Homogeneidade**: $P(N_{(i+1)} = y \mid N_{(i)} = x) = P(N_{(1)} = y \mid N_{(0)} = x)$, que significa que as probabilidades de transição entre os estados não mudam ao longo do tempo.
- **Irredutibilidade**: todos estados comunicam entre si
- **Parâmetro i contínuo**: o conjunto de índices de i é contínuo, $T = [0, \infty)$.

No contexto de filas, a homogeneidade significa que as probabilidades de mudança de quantidade de clientes no sistema não muda ao longo do tempo. A irredutibilidade significa que o sistema pode mudar para qualquer quantidade de clientes. E o parâmetro contínuo significa que os intervalos de tempo são contínuos. Como resultado destas suposições, tem-se que $N_{(i)}$ é um processo de nascimento e morte.

De acordo com Smith (2018), denota-se os conceitos de nascimento e morte já vistos anteriormente, no contexto da teoria de filas:

- **Nascimento** é a chegada de um novo cliente no sistema de filas.

- **Morte** é a saída de um cliente atendido.

Assim como as taxas de chegada, serviço, nascimento e morte.

(i) A **taxa de chegada** $P(N_{(s+t)} = n+1 \mid N_{(s)} = n) = \lambda t$ é a probabilidade de entrar um novo cliente no sistema em um intervalo t de tempo.

(ii) A **taxa de saída** $P(N_{(s+t)} = n-1 \mid N_{(s)} = n) = \mu t$ é a probabilidade de sair um cliente do sistema em um intervalo t de tempo.

Define-se $A = \{A_{(i)}, i \in T\}$, como um processo estocástico, tal que, para cada $i \in T$, $A_{(i)}$ é uma variável aleatória que representa o total de pessoas que chegaram no sistema até o tempo i. Além disso, é possível notar que a variável $A_{(i)}$ é um processo de contagem, já que $A_{(i)}$ também é um valor inteiro não negativo e para $s < i$, $A_{(s)} < A_{(i)}$ e $A_{(s)} - A_{(i)}$ é o número de eventos que ocorreram no intervalo $(s, i]$.

Como será considerado também que: o sistema de filas sempre começa vazio $A_{(0)} = 0$, o processo com incrementos independentes e o número de eventos em qualquer intervalo de comprimento i tem distribuição Poisson com média λi

$$P(A_{(t+s)} - A_{(s)} = x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}$$

Logo, $A_{(i)}$ é um Processo Poisson.

A principal diferença entre os processos N e A é que o processo N descreve a quantidade de pessoas presentes no sistema em um certo intervalo de tempo, enquanto o processo A descreve o total de pessoas que chegaram no sistema até um certo intervalo.

Além desses processos, define-se $D_{(t)} = A_{(t)} - N_{(t)} =$ número de pessoas que saíram do sistema até o instante t.

2.2.2 Características das filas

Existem vários tipos de sistemas de filas. Antes de analisá-los, é necessário definir bem suas características. As principais características de um sistema de filas são:

- **Padrão de entrada ou chegada de clientes**: a forma como as chegadas ocorrem. É especificada pelo intervalo entre duas chegadas consecutivas. A distribuição de tempo é considerada exponencial.

- **Padrão de serviço**: é a maneira em que o serviço é feito. É especificado pelo tempo que se leva para completar um serviço, com distribuição exponencial.
- **Número de servidores ou canais de serviço**: um sistema pode ter um único servidor ou um número de servidores paralelos. Um cliente que chegar e encontrar mais de um servidor pode escolher aleatoriamente qualquer um deles.
- **Capacidade do sistema**: um sistema pode ter uma capacidade infinita, isto é, a fila pode crescer a qualquer extensão ou finita. Se for finita, ela será especificada pelo número de lugares disponíveis para a fila.
- **Disciplina da fila**: indica a maneira em que cada unidade é atendida. Por exemplo, *first in - first out (FIFO)*, sistema em que os usuários são atendidos na ordem das chegadas que é a disciplina mais comum.

Um exemplo de um sistema de filas é o atendimento de emergência de um hospital, no qual é possível supor que as chegadas ocorrem de acordo com uma distribuição exponencial (padrão de entrada), o tempo para o atendimento ser feito tem distribuição exponencial (padrão de serviço), existem vários possíveis médicos para realizar o atendimento (número de servidores), a fila não é limitada (capacidade do sistema) e os atendimentos ocorrem conforme a chegada dos pacientes (disciplina da fila). Um exemplo de sistema de filas que não ocorre conforme a chegada dos clientes é o *First Expire, First Out*, que significa que o primeiro a ter uma certa validade, será atendido primeiro. O que pode acontecer com produtos perecíveis no supermercado.

2.2.3 Modelos de filas

As nomenclaturas dos modelos seguem a notação de Kendall (1953).A primeira letra refere-se a como as chegadas ocorrem, a segunda letra aos atendimentos, o número a seguir mostra a quantidade de postos de atendimentos. Em seguida, a capacidade do sistema (ilimitada ou limitada). E, por último, em que ordem os atendimentos são feitos. Como neste trabalho apenas os sistemas mais simples foram feitos, apenas sistemas com *M/M* serão vistos, que tanto as chegadas quanto os atendimentos são processos markovianos com distribuição exponencial.

2.2.3.1 M/M/1/∞/FIFO

Este é o modelo mais simples de filas, pode ser representado como M/M/1. Um exemplo de M/M/1 é uma loja com um único caixa e sem preferencialidade. Nesse

sistema, a primeira pessoa que entra na fila será a primeira a ser atendida, apenas uma pessoa por vez pode realizar o pagamento e não há um limite definido para quantas pessoas aguardam na fila. Tem as seguintes características:

- *FIFO* - os usuários são atendidos na ordem das chegadas;
- Tempos de chegada seguem distribuição exponencial de parâmetro λ;
- Tempos de atendimento seguem distribuição exponencial de parâmetro μ;
- Um único posto de atendimento;
- Não há um limite de tamanho da fila.

2.2.3.2 M/M/1/*k*/FIFO

Pode ser representado como M/M/1/k. Suas características são iguais a do M/M/1 com exceção do tamanho da fila, que se limita a um k. Portanto, suas características são:

- Atendimento conforme a chegada na fila;
- Tempos de chegada seguem distribuição exponencial de parâmetro λ;
- Tempos de atendimento seguem distribuição exponencial de parâmetro μ;
- Um único posto de atendimento;
- Há um limite de tamanho k para a fila.

2.2.3.3 M/M/c/∞/FIFO

Este modelo tem c postos de atendimentos ao invés de um único, como o M/M/1. Pode ser exemplificado por um supermercado com vários caixas e uma única fila.

- Atendimento conforme a chegada na fila;
- Tempos de chegada seguem distribuição exponencial de parâmetro λ;
- Tempos de atendimento seguem distribuição exponencial de parâmetro μ;
- Tem c postos de atendimentos;
- Não há um limite de tamanho da fila.

2.2.3.4 M/M/c/k/FIFO

Este modelo é similar ao M/M/c, mas existe um limite para a quantidade de pessoas na fila. Por exemplo, um supermercado com inúmeros caixas, uma única fila que apenas pode chegar ao tamanho k.

- Atendimento conforme a chegada na fila;
- Tempos de chegada seguem distribuição exponencial de parâmetro λ;
- Tempos de atendimento seguem distribuição exponencial de parâmetro μ;
- c postos de atendimentos;
- Fila de tamanho máximo k.

2.2.3.5 Caso Particular: M/M/c/c/FIFO

Neste modelo, não há formação de filas, o cliente só entra no sistema se houver um posto de atendimento livre.

- Atendimento conforme a chegada na fila
- Tempos de chegada seguem distribuição exponencial de parâmetro λ
- Tempos de atendimento seguem distribuição exponencial de parâmetro μ
- c postos de atendimentos
- Fila de tamanho máximo c.

2.2.4 Medidas de desempenho

Agora serão apresentadas medidas comumente utilizadas para analisar e comparar os modelos, encontradas em Mendonça (2014). As medidas são importantes para o estudo e comparação de sistemas. Neste trabalho, também será útil para verificar se o método de simulação dos sistemas está condizente com a teoria de filas. Podem-se citar:

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

- Probabilidade de se ter mais de k elementos no sistema ($P(N \geq k)$) ;
- Função de distribuição acumulada do tempo de espera na fila ($W_{q(t)}$);
- Função de densidade do tempo de espera na fila ($w_{q(t)}$);
- Função de distribuição acumulada do tempo de permanência no sistema ($W(t)$).

Os parâmetros são valores feitos a partir de λ e μ, e são usados na construção das medidas de desempenho:

- ρ é a taxa de ocupação utilização do sistema;
- P_0 é a probabilidade de haver 0 usuários no sistema;
- P_n é a probabilidade de haver n usuários no sistema;
- P_c é a probabilidade de usuários rejeitados pela limitação de c usuários na fila.

Nem sempre são conhecidas todas as medidas de desempenho para todos os sistemas apresentados. Vejamos aquelas conhecidas.

2.2.4.1 M/M/1/∞/FIFO

O modelo M/M/1 é o sistema mais simples de todos. Esta seção contém a apresentação e algumas demonstrações de suas medidas de desempenho. Para os modelos seguintes, serão apresentados apenas os resultados, dos quais as demonstrações se encontram na Seção 2.6 em Mendonça (2014).

Parâmetros

As taxas de chegada e de saída são constantes e dadas por:

$$\lambda_n = \lambda \ \forall n \geq 0 \text{ e } \mu_n = \mu \ \forall n \geq 1.$$

Além disso,

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

. Ou seja, as probabilidades de mudança de estados não mudam dependendo do instante t. Para este processo, obtém-se os seguintes resultados, demonstrados por Mendonça (2014):

$$P_n = \frac{\lambda^n}{\mu^n} P_0, \ \forall n \geq 1 \text{ e } P_0 = \left[\sum_{n=0}^{\infty} (\frac{\lambda}{\mu})^n \right], \tag{2.4}$$

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

$$P_0 = 1 - \frac{\lambda}{\mu}. \tag{2.5}$$

O parâmetro ρ é denominado como taxa de ocupação/utilização do sistema, que é dado por:

$$\rho = \frac{\lambda}{\mu} \tag{2.6}$$

que substituindo (2.5) e (2.6) em (2.4) leva a:

$$P_n = \rho^n(1-\rho),\ \forall n \geq 0 \tag{2.7}$$

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

Ainda considerando o sistema no regime estacionário, seja T_q a variável aleatória contínua que representa o tempo que um usuário permanece na fila. Esse tempo depende de quantas pessoas estão na frente e de quanto tempo cada atendimento terá. É possível identificar dois casos na chegada de cada usuário no sistema:

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 $t > 0$,

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

Definindo $T_{(n)}$ como a variável aleatória contínua que representa a soma dos tempos de atendimento de n usuários consecutivos. Os tempos de serviços são independentes e exponencialmente distribuídos com taxa μ. Como consequência, $T_{(n)}$ segue uma distribuição

de Erlang de parâmetros n e μ. Assim, $\forall t \geq 0$,

$$\begin{aligned} W_q(t) &= W_q(0) + \sum_{n=1}^{\infty} P[(n \text{ usuários no sistema } \cap (T_{(n)} \leq t))] && (2.8) \\ &= P_0 + \sum_{n=1}^{\infty} P_n P[T_{(n)} \leq t | n \text{ usuários no sistema}] && (2.9) \\ &= (1-\rho) + \sum_{n=1}^{\infty} [\rho^n (1-\rho)] \left(\int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} \right) dx && (2.10) \\ &= (1-\rho) + \rho(1-\rho)\mu \int_0^t e^{-(\mu-\lambda)x} dx. && (2.11) \end{aligned}$$

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

$$\begin{aligned} W_q(t) &= (1-\rho) + \rho - \rho e^{-(\mu-\lambda)t} && (2.12) \\ &= 1 - \rho e^{-(\mu-\lambda)t}. && (2.13) \end{aligned}$$

Função de densidade do tempo de espera na fila ($w_{(q)}(t)$)

Derivando-se a sua função de distribuição acumulada, obtém-se a função de densidade:

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

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

Assim como foi obtida $W_q(t)$, pode-se obter a distribuição acumulada do tempo T de permanência no sistema, $W(t)$:

$$W(t) = 1 - e^{-\mu(1-\rho)t} \; \forall t \geq 0. \tag{2.14}$$

Para os demais sistemas de fila os resultados encontram-se nos Apêndices 1 e 2, que foram retirados de Mendonça (2014). **Tempo médio de espera na fila (W_q)**

Seja T_q a variável aleatória contínua que representa o tempo que um usuário permanece na fila. Isso depende do número de unidades que estão na frente e do tempo dos atendimentos. 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 calculado observando-se que o tempo médio que um usuário permanece no sistema é igual à soma do tempo médio de espera na fila com o tempo médio de atendimento:

$$W = W_q + \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)} + \frac{1}{\mu} = \frac{1}{\mu - \lambda} \tag{2.15}$$

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:

$$N_q = \begin{cases} N - 1, & \forall N \geq 1, \\ 0, & N = 0, \end{cases} \tag{2.16}$$

de onde

$$L_q = E[N_q] = \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=1}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n = L - 1 + P_0 \tag{2.17}$$

$$= \frac{\rho}{1-\rho} - 1 + 1 - \rho = \frac{\rho - \rho + \rho^2}{(1-\rho)} = \frac{\rho^2}{(1-\rho)} \tag{2.18}$$

Número médio de usuários no sistema (L)

A taxa média de ingressos no sistema é representado por $E[\Lambda]$. E o número médio de usuários num sistema é dado por

$$L = E[\Lambda]W.$$

Além disso, há outras relações, como:

$$W = W_q + E[S]$$

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

$$W_q = \frac{L_q}{E[\Lambda]} \text{ e } L_q = L - E[\Lambda]E[S].$$

Assim, para este modelo, tem-se $E[\Lambda]$ e $E[S] = \frac{1}{\mu}$, portanto:

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

Probabilidade de se ter mais do que k elementos no sistema ($P(N \geq k)$)

$$P(N \geq k) = \sum_{n=k}^{\infty} P_n = \sum_{n=k}^{\infty} \rho^n(1-\rho) = (1-\rho)\sum_{i=0}^{\infty} \rho^{k+1} \quad (2.19)$$

$$= (1-\rho)\rho^k \sum_{i=0}^{\infty} \rho_i = (1-\rho)\rho^k \frac{1}{(1-\rho)}, \quad (2.20)$$

de onde segue que

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

2.2.4.2 M/M/1/*k*/FIFO

Parâmetros

As taxas de chegada e saída:

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

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

A taxa de ocupação de utilização do sistema

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

As probabilidades de haver 0 e n usuários no sistema:

$$P_0 = \begin{cases} \frac{1}{k+1}, & \rho = 1 \\ \frac{1-\rho}{1-\rho^{k+1}}, & \rho \neq 1 \end{cases} \quad P_n = \begin{cases} \frac{1}{k+1}, & \rho = 1 \\ \frac{(1-\rho)\rho^n}{1-\rho^{k+1}}, & \rho \neq 1 \end{cases}$$

Tempo médio de permanência no sistema (W)

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

Tempo médio de espera na fila (W_q)

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

Número médio de usuários no sistema (L)

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

Número médio de usuários na fila (L_q)

$$L_q = L - 1 + P_0$$

Probabilidade de se ter mais do que k elementos no sistema ($P(N \geq k)$)

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

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

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

$$W_q(t) = 1 - \sum_{n=0}^{n-2} q_{n+1} \sum_{i=0}^{n} \frac{(\mu t)^i}{i!} e^{-\mu t}$$

2.2.4.3 M/M/c/∞/FIFO

Parâmetros

As taxas de chegada e saída:

$$\lambda_n = \lambda\ \forall n \geq 0\ \ \mu_n = \begin{cases} n\mu, & \text{se } 1 \leq n < c, \\ c\mu, & \text{se } n \geq c. \end{cases}$$

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

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

A probabilidade de haver 0 e n usuários no sistema:

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{cr^c}{c!(c-r)} \right)^{-1} \quad P_n = \begin{cases} P_0 \frac{r^n}{n!}, & \text{se } 1 \leq n < c, \\ P_0 \frac{r^n}{c^{n-c}c!}, & \text{se } n \geq c \end{cases}$$

Tempo médio de permanência no sistema (W)

$$W = \frac{1}{\mu} + \left[\frac{r^c \mu}{(c-1)!(c\mu-\lambda)^2}\right] P_0$$

Tempo médio de espera na fila (W_q)

$$W_q = \frac{r^c \mu}{(c-1)!(c\mu-\lambda)^2} P_0$$

Número médio de usuários na fila (L_q)

$$L_q = \frac{P_0 c r^{c+1}}{c!(c-r)^2}$$

Número médio de usuários no sistema (L)

$$L = r + \left[\frac{r^{c+1} c}{c!(c-r)^2}\right] P_0$$

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

$$W_q(t) = 1 - P_0 \frac{r^c}{c!(1-\rho)} e^{-(c\mu-\lambda)t}$$

2.2.4.4 M/M/c/k/FIFO

Parâmetros

As taxas de chegada e saída são dadas, respectivamente, por

$$\lambda_n = \begin{cases} \lambda, & 0 \leq n < k, \\ 0, & n \geq k \end{cases} \text{ e } \mu_n = \begin{cases} n\mu, & 1 \leq n < c, \\ c\mu, & c \leq n \leq k. \end{cases}$$

A probabilidade de haver n e 0 usuários no sistema

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

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

Tempo médio de permanência no sistema (W)

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

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

Tempo médio de espera na fila (W_q)

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

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

Número médio de usuários na fila (L_q)

$$L_q = \frac{P_0 r^{c+1}}{c!c} . \frac{([(\frac{r}{c}) - 1](k - c + 1)(\frac{r}{c})^{k-c} + 1 - (\frac{r}{c})^{k-c+1})}{(1 - (\frac{r}{c}))}$$

Número médio de usuários no sistema (L)

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

2.2.4.5 Caso Particular: M/M/c/c/FIFO

Parâmetros

As taxas de chegada e saída:

$$\lambda_n = \begin{cases} \lambda, & 0 \leq n < c, \\ 0, & n \geq c, \end{cases} \text{ e } \mu_n = n\mu,\ 1 \leq n \leq c.$$

As probabilidades de haver n e 0 usuários no sistema:

$$P_n = \frac{(\frac{r^n}{n!})}{\sum_{i=0}^{c} \frac{r^i}{i!}} .\ 0 \leq n \leq c \quad P_0 = \left[\sum_{n=0}^{c} \frac{r^n}{n!}\right]^{-1} \quad P_c = \frac{(\frac{r^c}{c!})}{\sum_{i=0}^{c} \frac{r^i}{i!}}$$

2.3 Simulação

Um dos objetivos deste trabalho é simular os sistemas de fila. A ideia é estimar os valores das medidas de desempenho a partir dos valores simulados.

2.3.1 Números Pseudo-aleatórios

Não há como gerar números verdadeiramente aleatórios computacionalmente por não ser um processo mecânico, como jogar uma moeda para observar a face que ficou para cima. Assim, utilizar-se-à métodos de gerar números pseudo-aleatórios que podem ser encontrados em alguns livros, como por exemplo (ROSS, 2014a).

Para gerar uma sequência de números pseudo-aleatórios primeiro é definido um valor

inicial X_0, chamado de semente, então os demais números desta sequência são calculados recursivamente, especificando-se positivos inteiros a, c e m, da seguinte forma:

$$X_{n+1} = (aX_n + c) \mod m,\ n \geq 0 \tag{2.21}$$

O que significa que $aX_n + c$ é dividido por m e o resto é usado como o valor de X_{n+1}. A quantidade X_n/m será levada como uma aproximação para a variável aleatória de uma distribuição uniforme no intervalo $(0, 1)$. Esse método é um dos mais simples e é chamado de gerador congruencial linear (LCG).

De acordo com Ross (2022), aconselha-se escolher alguns valores para a e m a fim de que a repetição da semente X_0 ocorra depois de muitos valores serem gerados. Para que isso aconteça, m deve ser um número primo grande que é suportado pela memória do computador. Critérios que os parâmetros devem satisfazer:

(i) Para qualquer semente inicial, a sequência resultado tem a aparência de ser uma sequência independente de variáveis aleatórias de distribuição $U(0, 1)$

(ii) Para qualquer semente inicial, o número de variáveis que podem ser geradas antes da repetição acontecer é grande

(iii) Os valores podem ser computados eficientemente pelo computador

Os valores $a = 7^5 = 16807$ e $m = (2^{35} - 31)$, que contemplam os critérios, foram sugeridos por Ross e serão os valores escolhidos pra as simulações.

Como exemplo, serão calculados os 10.000 termos desta sequência com $X_0 = 10$ e

$c = 0$:

$$\begin{aligned}
X_0 &= 10 \\
X_1 = 16807 \cdot 10 \mod (2^{35} - 31) &= 1585457046 \\
X_2 = 16807 \cdot 1585457046 \mod (2^{35} - 31) &= 17979865142 \\
X_3 = 16807 \cdot 17979865142 \mod (2^{35} - 31) &= 28054758115 \\
X_4 = 16807 \cdot 28054758115 \mod (2^{35} - 31) &= 31990279331 \\
X_5 = 16807 \cdot 31990279331 \mod (2^{35} - 31) &= 33798990691 \\
X_6 = 16807 \cdot 33798990691 \mod (2^{35} - 31) &= 24442373151 \\
X_7 = 16807 \cdot 24442373151 \mod (2^{35} - 31) &= 32293881283 \\
X_8 = 16807 \cdot 32293881283 \mod (2^{35} - 31) &= 16835985727 \\
&\vdots \\
X_{9999} = 16807 \cdot 17045897409 \mod (2^{35} - 31) &= 33259506405 \\
X_{10000} = 16807 \cdot 33259506405 \mod (2^{35} - 31) &= 28300899321
\end{aligned}$$

Fazendo $u_i = X_i/m$ para cada i e arredondando para 4 casas decimais, obtém-se os números:

u_i	**Valor**
u_1	0,0461
u_2	0,5233
u_3	0,8165
u_4	0,9310
u_5	0,9837
u_6	0,7114
u_7	0,9399
u_8	0,4900
$\vdots$	$\vdots$
u_{9999}	0,9680
u_{10000}	0,8237

Dessa forma, tem-se a geração de 10.000 termos de uma sequência pseudo-aleatória, que será considerada como uma geração de uma amostra aleatória com distribuição $U(0, 1)$, importante para a geração de amostras de outras distribuições.

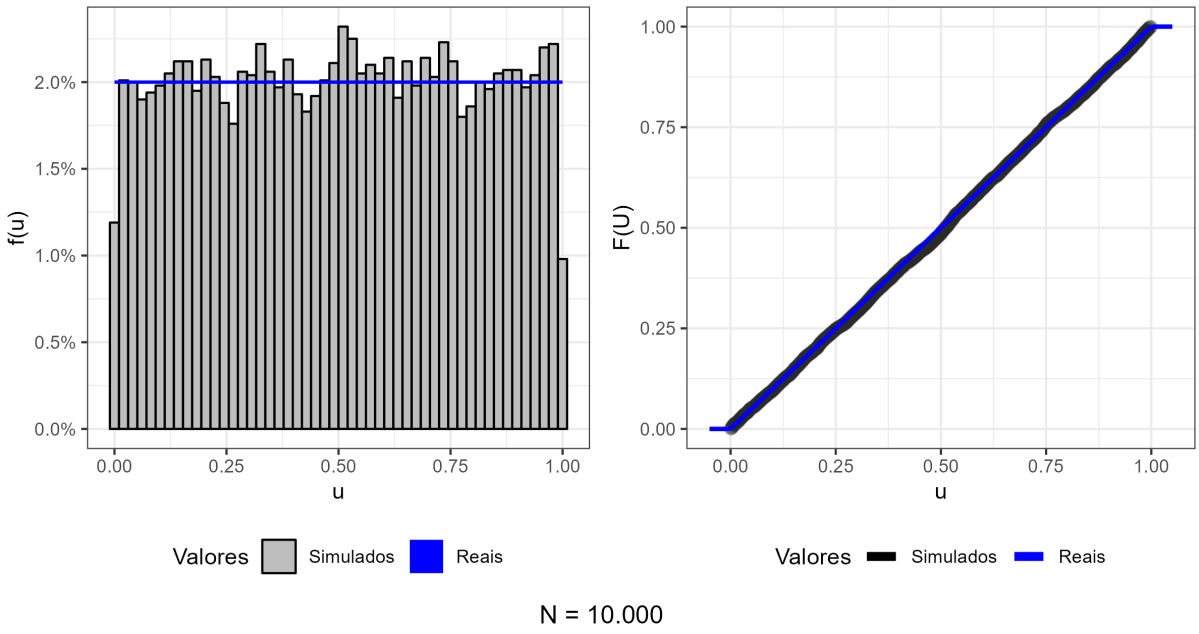

Figura 1: Comparação entre a distribuição uniforme (0,1) e os valores gerados

2.3.2 Como gerar uma amostra de exp(λ)

A partir da geração da uniforme $(0, 1)$, torna-se possível gerar amostras de outras distribuições. O método usado para gerar uma amostra pseudoaleatória da distribuição exponencial de parâmetro λ será o da Transformada Inversa, usado para gerar variáveis aleatórias de distribuições contínuas com função de distribuição acumulada conhecida e inversível.

O método consiste em utilizar a função de distribuição acumulada inversa da distribuição de interesse F^{-1} e o valor u da $U(0, 1)$ gerado pelo método LCG nessa função.

Proposição 2.1 *U é uma variável aleatória uniforme $U(0, 1)$. Para qualquer distribuição contínua F se definirmos a variável aleatória Y por*

$$Y = F^{-1}(U)$$

então a variável aleatória Y tem uma função de distribuição F. [$F^{-1}(u)$ é definida para ser igual ao valor y, para o qual $F(y) = u$]

Assim, como a função de distribuição da exponencial é contínua, pode-se utilizar esse método para gerar amostras aleatórias.

A F de uma distribuição exponencial é $1 - e^{-y}$, então $F^{-1}(u)$ é o valor de y em que

$$1 - e^{-y} = u$$

ou

$$y = -\log(1 - u)$$

Já que U é uma variável uniforme $(0, 1)$, então

$$F^{-1}(U) = -\log(1 - U)$$

é uma variável aleatória exponencialmente distribuída. Além disso, como $1 - U$ também é uniformemente distribuído no intervalo $(0, 1)$, segue que $-\log(1 - U)$ é exponencialmente distribuído com média 1, e $-c\log(1 - U)$ é exponencialmente distribuído com média c.

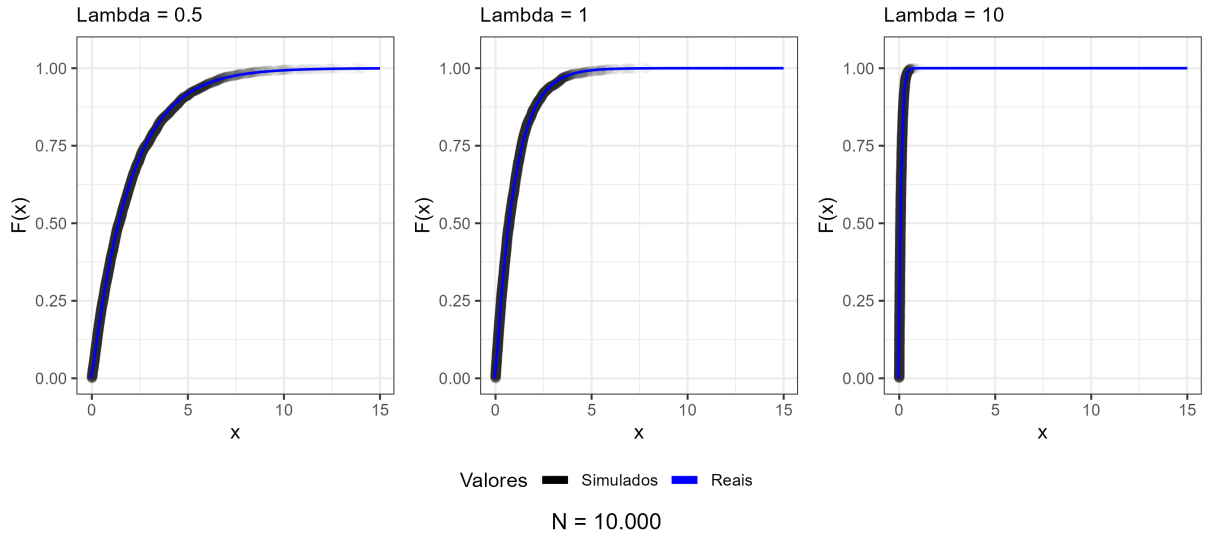

Figura 2: Distribuição acumulada exponencial real e simulada

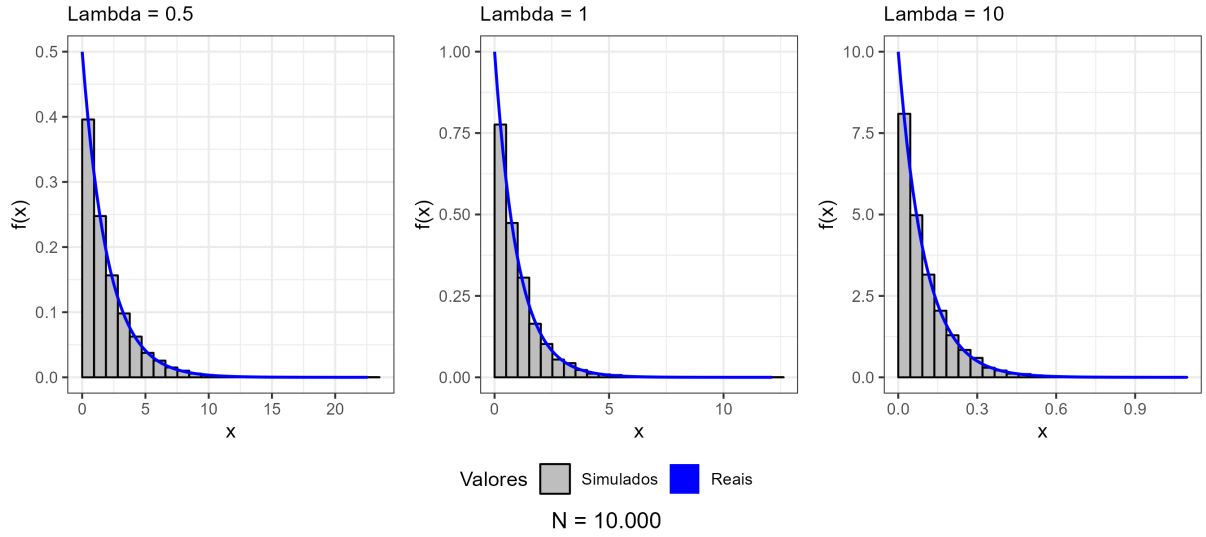

Figura 3: Distribuição da densidade exponencial real e simulada

Os gráficos da Figura 2 comparam os valores reais e simulados da distribuição acumulada da exponencial. Nota-se que para diferentes valores de λ, os valores simulados se aproximam aos valores reais. A Figura 3 realiza algo semelhante para a distribuição de densidade, e a distribuição simulada também converge.

3 Resultados

O software utilizado para as simulações é R Core Team (2022). Serão simulados diversos sistemas com taxas de chegada e de atendimento diferentes, cada sistema será simulado 1 vez considerando uma janela de tempo de 25000 minutos. As medidas de desempenho serão estimadas a partir dos valores computados nas simulações. Por exemplo, em cada simulação, para estimar o tempo médio de permanência no sistema (W), será computado o total de tempo de permanência no sistema de todos clientes e isso será dividido pelo número total de clientes. Para checar se a medida está convergindo para o valor teórico do sistema, essa medida será estimada a cada cliente novo no sistema ou a cada mudança de estado do sistema, dependendo da medida calculada. Assim, haverá uma estimativa final para cada medida em todas as simulações realizadas. A Tabela 1 e 2 mostram as combinações de parâmetros simuladas neste trabalho:

$\boldsymbol{\lambda}$	$\boldsymbol{\mu}$
0.50	1.00
0.75	1.00
0.50	2.00
0.75	2.00
1.00	2.00

Tabela 1: M/M/1/∞

$\boldsymbol{\lambda}$	$\boldsymbol{\mu}$	**k**
0.50	1.00	5
0.75	1.00	5
0.50	1.00	10
0.75	1.00	10

Tabela 2: M/M/1/k

É interessante lembrar que λ é o parâmetro dos tempos de chegada entre os clientes, $A_{(i)} \backsim Exp(\lambda)$. Logo, $E(A) = \frac{1}{\lambda}$, ou seja, quanto menor for λ, maiores serão os intervalos de chegada. Analogamente, μ é o parâmetro dos tempos de atendimento dos clientes, $D_{(i)} \backsim Exp(\mu)$. Assim, quanto menor for μ, maiores os tempos para os clientes serem atendidos. Como consequência, quanto mais próximo λ for de μ, mais ocupado o sistema será, pois os clientes chegarão com uma frequência próxima da frequência de que estão sendo atendidos. Além disso, para as filas não ter um valor infinido, $\lambda < \mu$, visto que

$\rho = \frac{\lambda}{\mu} \leq 1$ e é a taxa de utilização do sistema ou a probabilidade do sistema estar ocupado.

3.1 M/M/1/∞/FIFO

Este é o sistema mais simples. Os clientes são atendidos conforme a chegada, há apenas um posto de atendimento e não há um limite para o tamanho da fila. Os parâmetros e as medidas de desempenho podem ser encontradas, respectivamente, nos Apêndices 1 e 2.

Tabela 3: Parâmetros e Medidas Teóricas

$\boldsymbol{\mu}$	$\boldsymbol{\lambda}$	$\boldsymbol{\rho}$	$\boldsymbol{W}$	$\boldsymbol{W_q}$	$\boldsymbol{L}$	$\boldsymbol{L_q}$
1.000	0.500	0.500	2.000	1.000	1.000	0.500
1.000	0.750	0.750	4.000	3.000	3.000	2.250
2.000	0.500	0.250	0.667	0.167	0.333	0.083
2.000	0.750	0.375	0.800	0.300	0.600	0.225
2.000	1.000	0.500	1.000	0.500	1.000	0.500

A Tabela 3 mostra os parâmetros e medidas de desempenho teóricos.

3.1.1 Pseudocódigo

1. definir
    - tempo: janela de tempo que o sistema ficará aberto
    - λ: parâmetro para a distribuição exponencial dos tempos de chegada entre os clientes
    - μ: parâmetro para a distribuição exponencial dos tempos de atendimento de cada cliente
    - k: parâmetro que limita o tamanho das filas
2. iniciar vetor chegadas *(tempos de chegadas entre os clientes)*
3. **enquanto** soma de chegadas < tempo
   chegada = exponencial λ
   concatenar chegada a chegadas
4. tirar o último item de chegadas que passa do tempo.
5. iniciar vetor HoraChegou *(hora que cada cliente chegou no sistema)*
6. HoraChegou[1] = chegadas[1]

7. **para** i de 2 a tamanho de chegadas
   HoraChegou[i] = HoraChegou[i-1] + Chegadas[i] *(Hora de chegadas)*
8. NumeroClientes = tamanho de chegadas
9. iniciar vetor Fila *(tamanho da fila para cada cliente)*
10. iniciar vetor HoraEntrou *(hora que cada cliente entrou em atendimento)*
11. iniciar vetor HoraSaiu *(hora que cada cliente saiu do sistema)*
12. iniciar vetor saturado *(vetor que guarda os clientes que não entraram no sistema porque a fila atingiu seu limite)*
13. fila = 0 *(controla o tamanho da fila)*
14. clienteatendido = 0 *(número de clientes atendidos até o instante t)*
15. ocorrencia = "chegada" *(controla qual a próxima ocorrência no sistema)*
16. atendimento = "sem cliente" *(controla o atendimento imediato ou a entrada na fila)*
17. t = HoraChegou[1] *(instante da próxima ocorrência)*
18. tproxsaida = ∞
19. iproxsaida = ∞
20. tproxchegada = HoraChegou[1]
21. iproxsaida = 1
22. **repetir**

    - **se** ocorrencia == "chegada"
      - **se** atendimento == "sem cliente"
        atendimento = "com cliente"
        concatenar t a HoraEntrou
        tempoatendimento = exponencial μ
        tproxsaida = t + tempoatendimento
        iproxsaida = iproxchegada
        iproxchegada = iproxchegada + 1
        tproxchegada = Horachegou[iproxchegada]
      - **caso contrário** *(atendimento == "com cliente")*
        fila = fila + 1
        concatenar fila a Fila
        iproxchegada = iproxchegada + 1
        tproxchegada = Horachegou[iproxchegada]

- **caso contrário** *(se ocorrencia == "saida")*
  concatenar t a HoraSaiu
  clienteatendido = clienteatendido + 1
  se clienteatendido == NumeroClientes **quebrar repetir**
  - **se** fila == 0
    tproxsaida = ∞
    iproxsaida = ∞
    atendimento = "sem cliente"
    concatenar fila a Fila
  - **caso contrário** *(se fila $\neq$ 0)*
    fila = fila - 1
    concatenar fila a Fila
    concatenar t a HoraEntrou
    tempoatendimento = exponencial μ
    tproxsaida = t + tempoatendimento
    iproxsaida = iproxsaida + 1
- **se** tproxsaida > tproxchegada
  t = tproxchegada
  ocorrencia = "chegada"
- **caso contrário** *(se tproxsaida < tproxchegada)*
  t = tproxsaida
  ocorrencia = "saida"
- voltar ao item 22

3.1.2 Simulação

A Figura 4 é o gráfico de todas as medidas de desempenho com $\mu = 1$ e $\lambda = 0.5$. Teoricamente, a média de tempo em que os clientes chegam é de 2 minutos e a média de atendimento é de 1 minuto. Nesse sistema, em geral, os clientes levam mais tempo para chegar do que para serem atendidos. A linha azul representa o valor teórico da medida. O restante dos gráficos com outros μ e λ estão no apêndice. As medidas de cada simulação, tempo médio de permanência no sistema, na fila e número médio de pessoas no sistema e na fila, convergem para a medida teórica

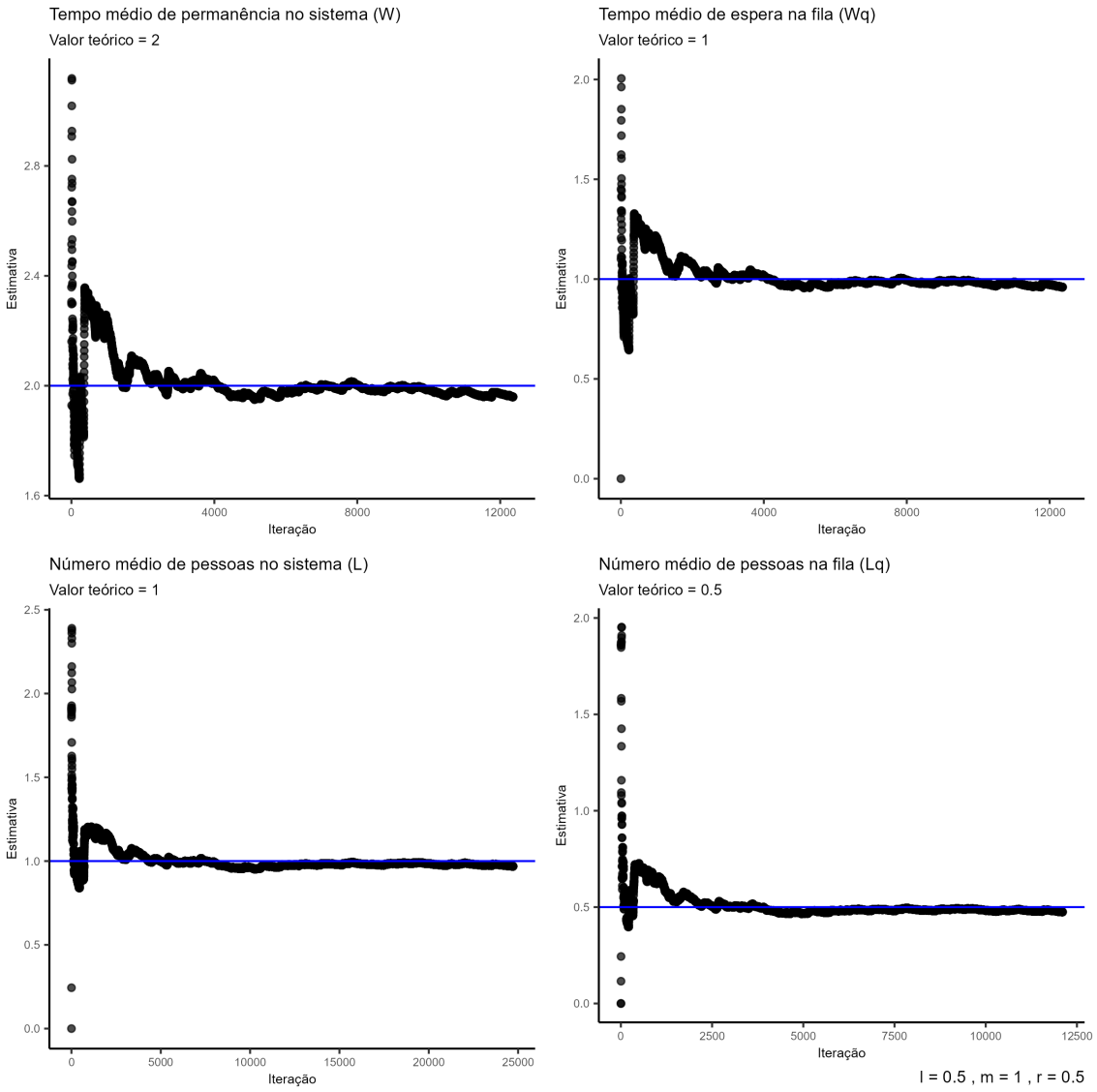

Figura 4: Medidas de desempenho

Nota-se que os valores teóricos para as medidas de desempenho aproximam-se aos seus respectivos valores teóricos. Os resultados para as outras combinações de λ e μ, apresentadas na Tabela 3, encontram-se no Apêndice 3.

3.2 M/M/1/k/FIFO

Este é um sistema muito similar ao M/M/1/∞/FIFO, mas a sua fila é limitada pelo parâmetro k. Os parâmetros e as medidas de desempenho podem ser encontradas, respectivamente, nos Apêndices 1 e 2.

Tabela 4: Parâmetros e Medidas Teóricas

μ	λ	k	ρ	W	W_q	L	L_q
1.000	0.500	5	0.500	1.838	0.838	0.905	0.413
1.000	0.500	10	0.500	1.990	0.990	0.995	0.495
1.000	0.750	5	0.750	2.444	1.444	0.701	1.005
1.000	0.750	10	0.750	3.403	2.403	2.515	1.776

A Tabela 4 mostra as combinações de parâmetros que foram simulados neste trabalho.

3.2.1 Pseudocódigo

1. definir
   - tempo: janela de tempo que o sistema ficará aberto
   - λ: parâmetro para a distribuição exponencial dos tempos de chegada entre os clientes
   - μ: parâmetro para a distribuição exponencial dos tempos de atendimento de cada cliente
2. iniciar vetor chegadas *(tempos de chegadas entre os clientes)*
3. **enquanto** soma de chegadas < tempo
   chegada = exponencial λ
   concatenar chegada a chegadas
4. tirar o último item de chegadas que passa do tempo.
5. iniciar vetor HoraChegou *(hora que cada cliente chegou no sistema)*
6. HoraChegou[1] = chegadas[1]
7. **para** i de 2 a tamanho de chegadas
   HoraChegou[i] = HoraChegou[i-1] + Chegadas[i] *(Hora de chegadas)*
8. NumeroClientes = tamanho de chegadas
9. iniciar vetor Fila *(tamanho da fila para cada cliente)*
10. iniciar vetor HoraEntrou *(hora que cada cliente entrou em atendimento)*
11. iniciar vetor HoraSaiu *(hora que cada cliente saiu do sistema)*
12. fila = 0 *(controla o tamanho da fila)*
13. clienteatendido = 0 *(número de clientes atendidos até o instante t)*
14. ocorrencia = "chegada" *(controla qual a próxima ocorrência no sistema)*

15. atendimento = "sem cliente" *(controla o atendimento imediato ou a entrada na fila)*
16. t = HoraChegou[1] *(instante da próxima ocorrência)*
17. tproxsaida = ∞
18. iproxsaida = ∞
19. tproxchegada = HoraChegou[1]
20. iproxsaida = 1
21. **repetir**

- **se** ocorrencia == "chegada"
  - **se** atendimento == "sem cliente"
    atendimento = "com cliente"
    concatenar t a HoraEntrou
    tempoatendimento = exponencial μ
    tproxsaida = t + tempoatendimento
    iproxsaida = iproxchegada
    iproxchegada = iproxchegada + 1
    tproxchegada = Horachegou[iproxchegada]
  - **caso contrário** *(atendimento == "com cliente")*
    - **se** a fila == k *(tamanho da fila chegou no limite)*
      concatenar iproxchegada a saturado
    - **caso contrário** *(se a fila ainda não chegou ao limite)*
      fila = fila + 1
  - concatenar fila a Fila
  - iproxchegada = iproxchegada + 1
  - tproxchegada = Horachegou[iproxchegada]
- **caso contrário** *(se ocorrencia == "saida")*
  proxsaida = clienteatendido + 1
  - **se** a proxsaida estiver no vetor saturado
    concatenar fila a Fila
    concatenar 0 a HoraSaiu
    clienteatendido = clienteatendido + 1
  - **caso contrário** concatenar fila a Fila
    concatenar t a HoraSaiu
    clienteatendido = clienteatendido + 1

- proxentrada = tamanho de HoraEntrou + 1
  - **se** proxentrada estiver no vetor saturado
  - **caso contrário**
    - **se** fila == 0

      tproxsaida = ∞

      iproxsaida = ∞

      atendimento = "sem cliente"

      concatenar fila a Fila
    - **caso contrário** *(se fila $\neq$ 0)*

      fila = fila - 1

      concatenar fila a Fila

      concatenar t a HoraEntrou

      tempoatendimento = exponencial μ

      tproxsaida = t + tempoatendimento

      iproxsaida = iproxsaida + 1
- **se** tproxsaida > tproxchegada

  t = tproxchegada

  ocorrencia = "chegada"
- **caso contrário** *(se tproxsaida < tproxchegada)*

  t = tproxsaida

  ocorrencia = "saida"
- voltar ao item 21

3.2.2 Simulação

A Figura 5 mostra a simulação para $\lambda = 0.5$, $\mu = 1$ e $k = 5$. Nesse sistema, em geral, os clientes levam mais tempo para chegar do que para serem atendidos. A linha azul representa o valor teórico da medida. O restante dos gráficos com outros μ,λ e k estão no apêndice.

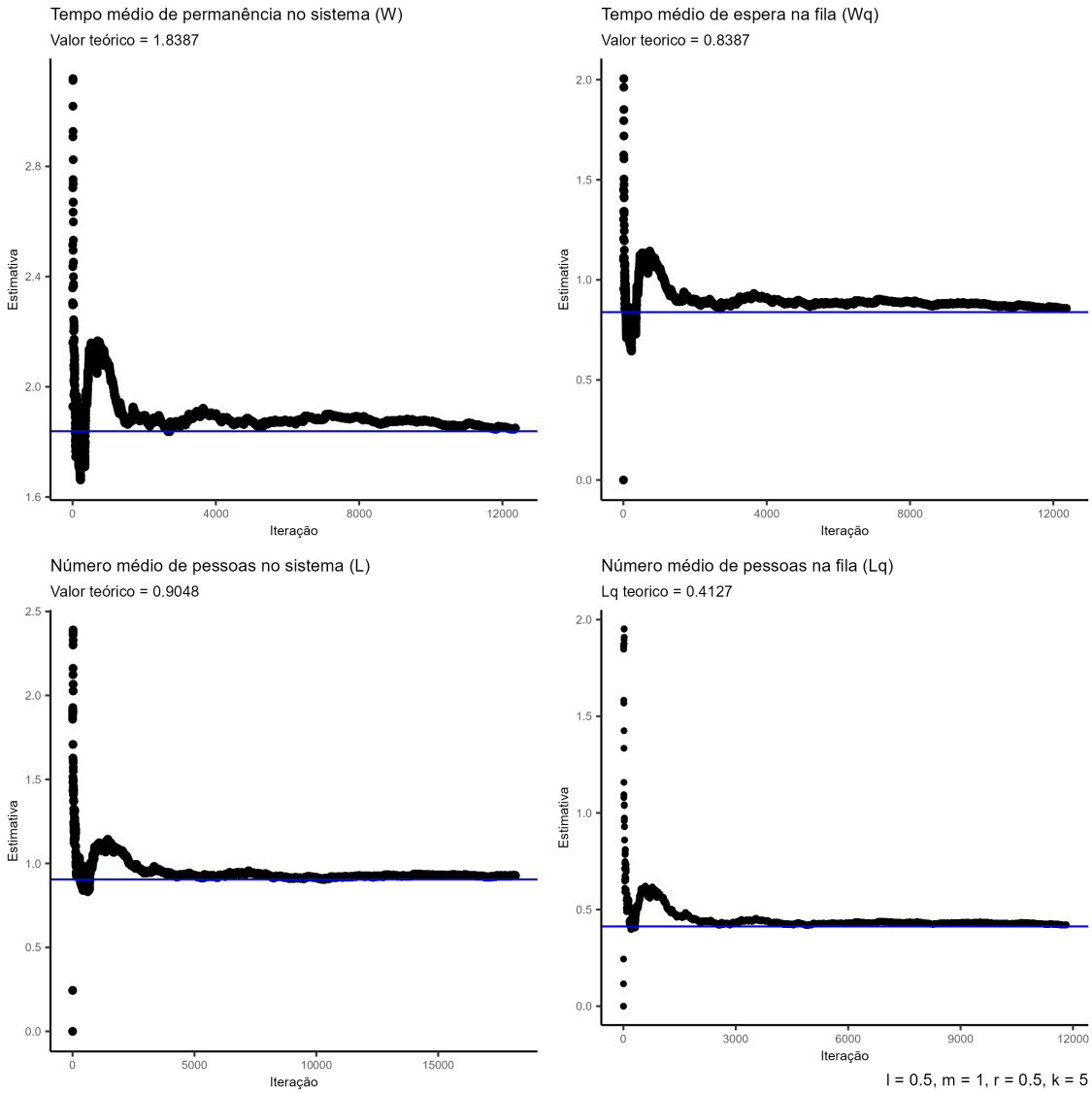

Figura 5: Medidas de desempenho

A Figura 5 mostra que os valores das medidas de desempenho simuladas convergem aos valores teóricos.

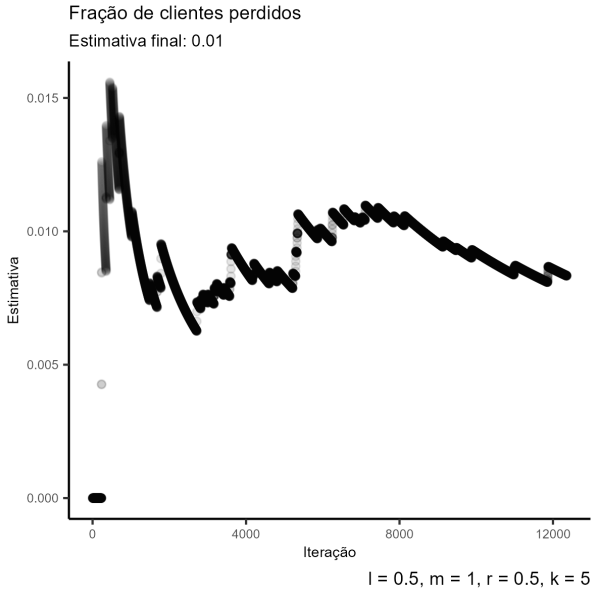

Figura 6: Sugestão de medida

A partir das simulações, pôde-se guardar quantas pessoas foram perdidas, pois chegaram ao sistema quando já estava lotado. Essa medida não tem um valor teórico para comparação, mas como os resultados anteriores convergiram, conclui-se que esse resumo da simulação é razoável para um estudo do sistema. A Figura 6 é a fração de pessoas perdidas para o mesmo sistema simulado da Figura 5, o valor estimado foi de apenas 1%, ou seja, apenas 1% dos clientes que chegaram ao sistema não foram atendidos.

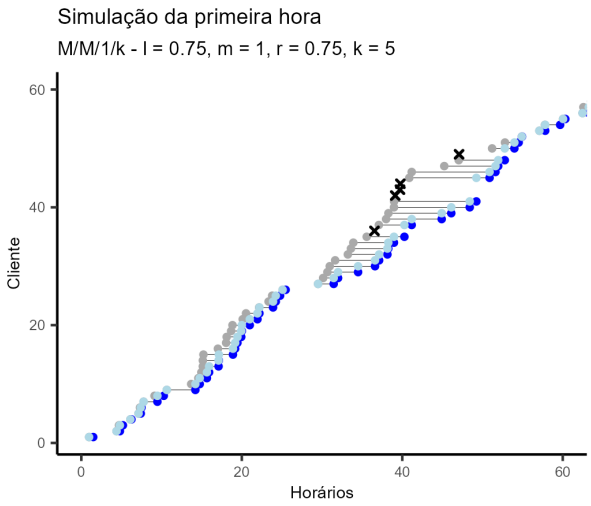

Figura 7: ponto cinza - chegadas/ azul claro - atendimento / azul escuro - saída / xis - cliente perdido

A Figura 7 mostra a primeira hora de uma simulação de um outro sistema, que é mais ocupado ($\lambda = 0.75$, $\mu = 1$ e $k = 5$). Na primeira hora, 4 clientes chegaram ao sistema e foram perdidos, pois estava lotado. Outros sistemas simulados se encontram no Apêndice 4.

4 Conclusões

Considerando a vasta área de aplicação que a Teoria de Filas tem, o trabalho teve como objetivo simular sistemas para poder analisar e entender o funcionamento, a fim de propor soluções.

Utilizando como base teórica, a Geração de Números Aleatórios e a Teoria de Filas, foram simulados tempos de chegada para cada cliente, e, passo a passo, seus tempos de atendimento em cadeia, ambos com distribuição exponencial.

Para obter todas as medidas de desempenho, foram coletados, separadamente, ao longo da simulação os tempos em que os estados do sistema e da fila mudavam e armazenadas em diferentes tabelas. Assim, foi possível calcular todas as medidas: tempo médio de permanência no sistema (W), número médio de pessoas na fila (W_q), número médio de pessoas no sistema (L) e número médio de pessoas na fila (L_q). Comparando essas medidas estimadas com as teóricas, concluiu-se que as simulações estavam obtendo resultados adequados.

Isto posto, para o sistema M/M/1/k, no qual um cliente é perdido quando este chega no sistema e a fila já está no tamanho k, foi sugerido uma nova medida que não apresenta valor teórico, a fração de clientes perdidos.

Foram simulados apenas dois sistemas, o mais simples, M/M/1/∞), e o M/M/1/k. Portanto, é possível sugerir para pesquisas futuras, simular outros sistemas, como M/M/c/∞ e M/M/c/k, que são sistemas com c postos de atendimento. E encontrar outras medidas sem valor teórico, como por exemplo a média de postos ocupados.

Referências

KENDALL, D. G. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, JSTOR, p. 338–354, 1953.

MENDONÇA, E. B. d. Teoria de filas markovianas e aplicações. *Orientadora: Divanilda Maia Esteves*, v. 63, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: ⟨https://www.R-project.org/⟩.

ROSS, S. M. *A first course in probability*. [S.l.]: Pearson, 2014.

ROSS, S. M. *Introduction to probability models*. [S.l.]: Academic press, 2014.

ROSS, S. M. *Simulation*. [S.l.]: academic press, 2022.

SMITH, J. M. *Introduction to queueing networks: theory and practice*. [S.l.]: Springer, 2018.

SMITH, J. S.; NELSON, B. L. Estimating and interpreting the waiting time for customers arriving to a non-stationary queueing system. In: IEEE. *2015 Winter Simulation Conference (WSC)*. [S.l.], 2015. p. 2610–2621.

WANG, S. *Discrete event simulation analysis on multiple class single server queueing system with abandonments and promotions*. Tese (Doutorado), 2019.

YAKIMOV, I. et al. The comparison of structured modeling and simulation modeling of queueing systems. In: SPRINGER. *International Conference on Information Technologies and Mathematical Modelling*. [S.l.], 2017. p. 256–267.

APÊNDICE 1 – Parâmetros dos modelos

Tabela 5: Parâmetros

Modelos	λ	μ	ρ	P_0	P_n	P_c
M/M/1	$\lambda_n = \lambda\ \forall n \geq 0$	$\mu_n = \mu\ \forall n \geq 1$	$\frac{\lambda}{\mu}$	$1 - \frac{\lambda}{\mu}$	$\frac{\lambda^n}{\mu^n} P_0\ \forall n \geq 1$	
M/M/1/k	$\lambda'_n = \begin{cases} \lambda, & 0 \leq n < k, \\ 0, & n \geq k \end{cases}$	$\mu_n = \mu\ \forall n \geq 1$	$\frac{\lambda}{\mu}$	$\begin{cases} \frac{1}{k+1}, & \rho = 1 \\ \frac{1-\rho}{1-\rho^{k+1}}, & \rho \neq 1 \end{cases}$	$\begin{cases} \frac{1}{k+1}, & \rho = 1 \\ \frac{(1-\rho)\rho^n}{1-\rho^{k+1}}, & \rho \neq 1 \end{cases}$	
M/M/c	$\lambda_n = \lambda\ \forall n \geq 0$	$\mu_n = \begin{cases} n\mu, & 1 \leq n < c, \\ c\mu, & n \geq c \end{cases}$	$\frac{\lambda}{c\mu}$	$(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{cr^c}{c!(c-r)})^{-1}$	$\begin{cases} P_0 \frac{r^n}{n!}, & 1 \leq n < c, \\ P_0 \frac{r^n}{c^{n-c}c!}, & n \geq c \end{cases}$	
M/M/c/k	$\lambda'_n = \begin{cases} \lambda, & 0 \leq n < k, \\ 0, & n \geq k \end{cases}$	$\mu_n = \begin{cases} n\mu, & 1 \leq n < c, \\ c\mu, & c \leq n \geq k \end{cases}$		$\begin{cases} [\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c(k-c+1)}{c!}]^{-1} & \frac{r}{c} = 1 \\ [\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c(1-[\frac{r}{c}]^{k-c+1})}{c!(1-\frac{r}{c})}] & \frac{r}{c} \neq 1 \end{cases}$	$\begin{cases} (\frac{r^n}{n!})P_0, & 1 \leq n \leq c-1 \\ (\frac{r^n}{c!c^{n-c}})P_0, & c \leq n \leq k. \end{cases}$	
M/M/c/c	$\lambda_n = \begin{cases} \lambda, & 0 \leq n < c \\ 0, & n \geq c \end{cases}$	$\mu_n = n\mu,\ 1 \leq n \leq c$	$[\sum_{n=0}^{c} \frac{r^n}{n!}]^{-1}$	$\left[\sum_{n=0}^{c} \frac{r^n}{n!}\right]^{-1}$	$\frac{(\frac{r^n}{n!})}{\sum_{i=0}^{c} \frac{r^i}{i!}}, 0 \leq n < c$	$\frac{(\frac{r^c}{c!})}{\sum_{i=0}^{c} \frac{r^i}{i!}}$

APÊNDICE 2 – Medidas de desempenho

Tabela 6: Medidas de Desempenho

Modelos	L_q	L	W	W_q	$P(N \geq k)$
M/M/1	$\frac{\rho^2}{(1-\rho)}$	$\frac{\rho^2}{1-\rho}$	$\frac{1}{\mu-\lambda}$	$\frac{\rho}{\mu-\lambda}$	ρ^k
M/M/1/k	$L-1+P_0$	$\frac{k}{2}$ se $\rho=1$	$\frac{L}{\lambda(1-P_k)}$	$\frac{L_q}{\lambda(1-P_k)}$	$\frac{k+1-k}{k+1}$ se $\rho=1$
		$\frac{\rho[1+k\rho^{k+1}-\rho^k(k+1)]}{(1-\rho)(1-\rho^{k+1})}$ se $\rho\neq 1$			$\rho^k\frac{1-\rho^{k+1-k}}{1-\rho^{k+1}}$ se $\rho\neq 1$
M/M/c	$\frac{P_0 c r^{c+1}}{c!(c-r)^2}$	$r+\left[\frac{r^{c+1}c}{c!(c-r)^2}\right]$	$\frac{1}{\mu}+\left[\frac{r^c\mu}{(c-1)!(c\mu-\lambda)^2}\right]P_0$	$\frac{r^c\mu}{(c-1)!(c\mu-\lambda)^2}$	
M/M/c/k	$\frac{P_0 r^{c+1}}{c!c}\frac{([\frac{r}{c}-1](k-c+1)(\frac{r}{c})^{k-c}+1-\frac{r}{c}^{k-c+1})}{(1-(\frac{r}{c}))}$	$L_q+c+\sum_{n=0}^{c-1}(n-c)P_n$	$\frac{L}{\lambda'}$	$\frac{L_q}{\lambda'}$	
M/M/c/c					

APÊNDICE 3 – Gráficos M/M/1/∞/*FIFO*

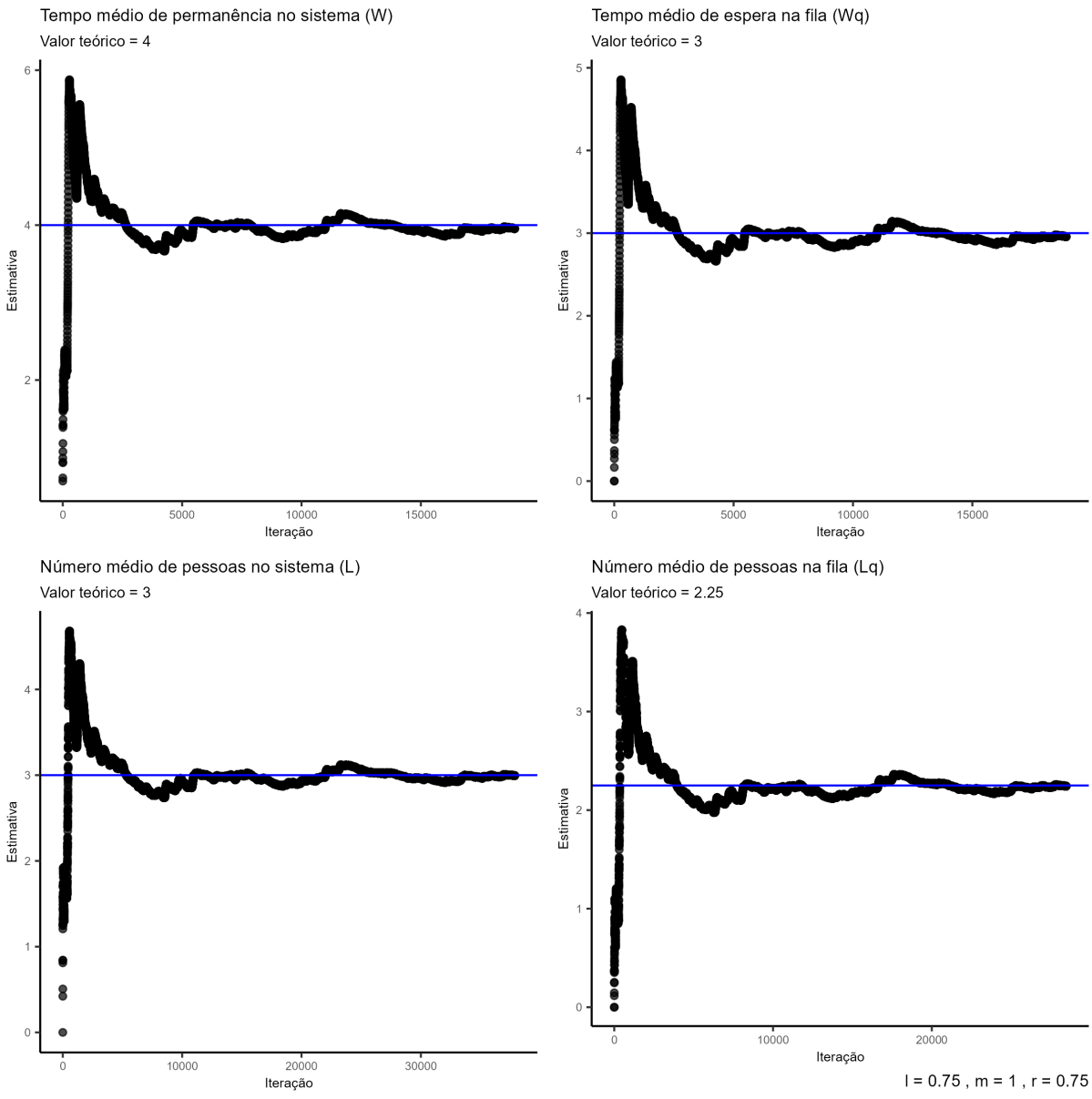

Figura 8: Sistema 1

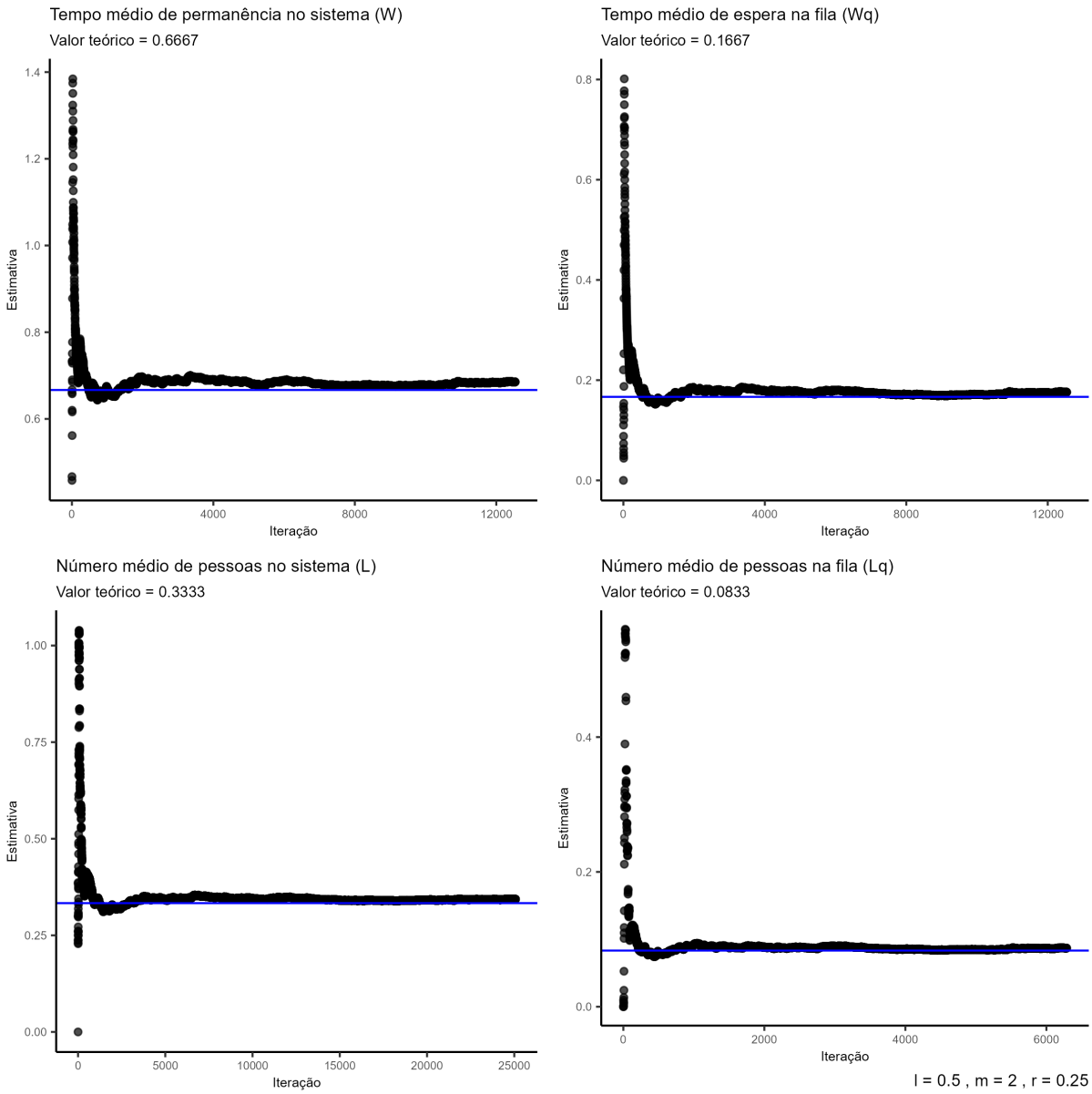

Figura 9: Sistema 2

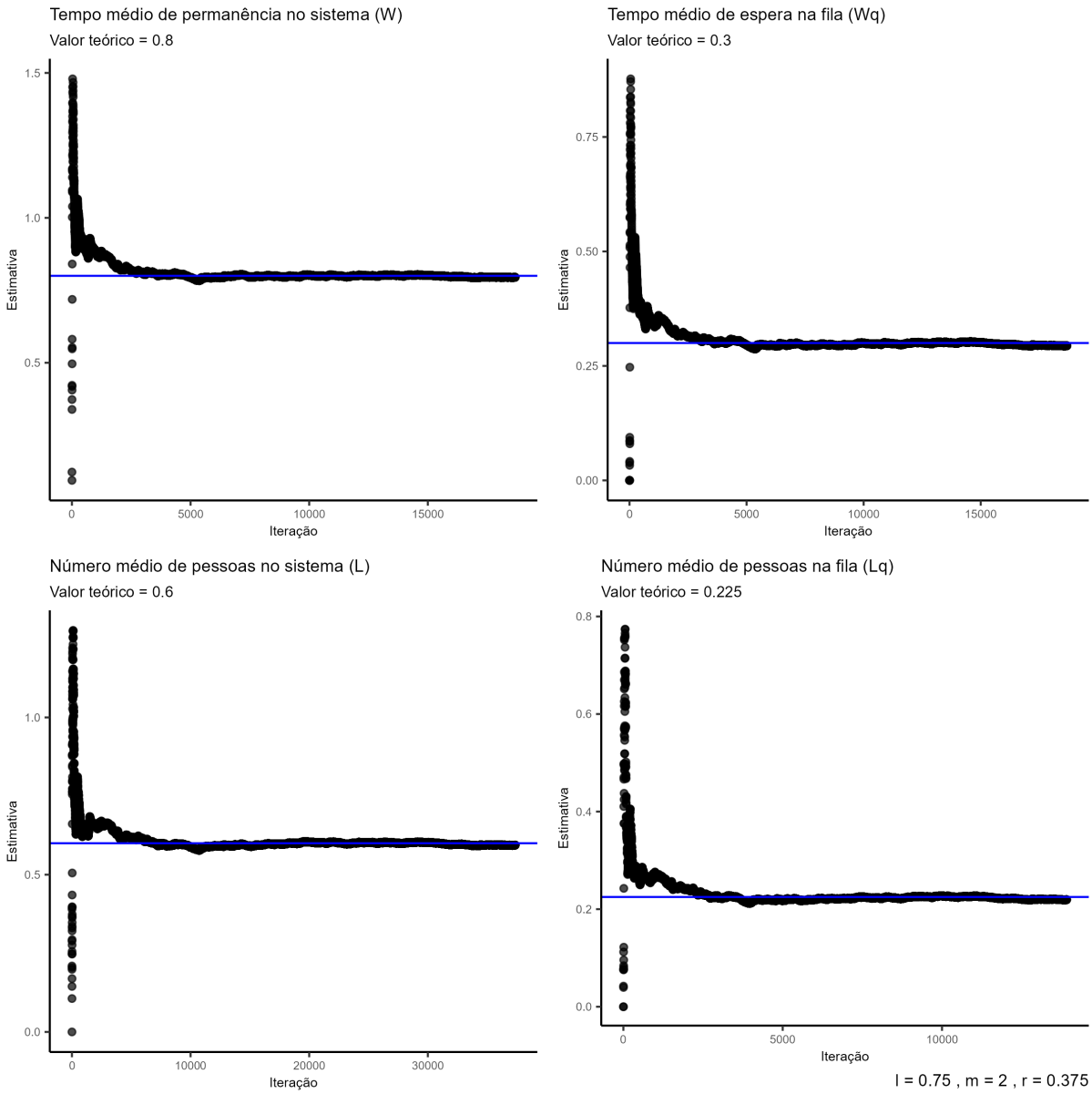

Figura 10: Sistema 3

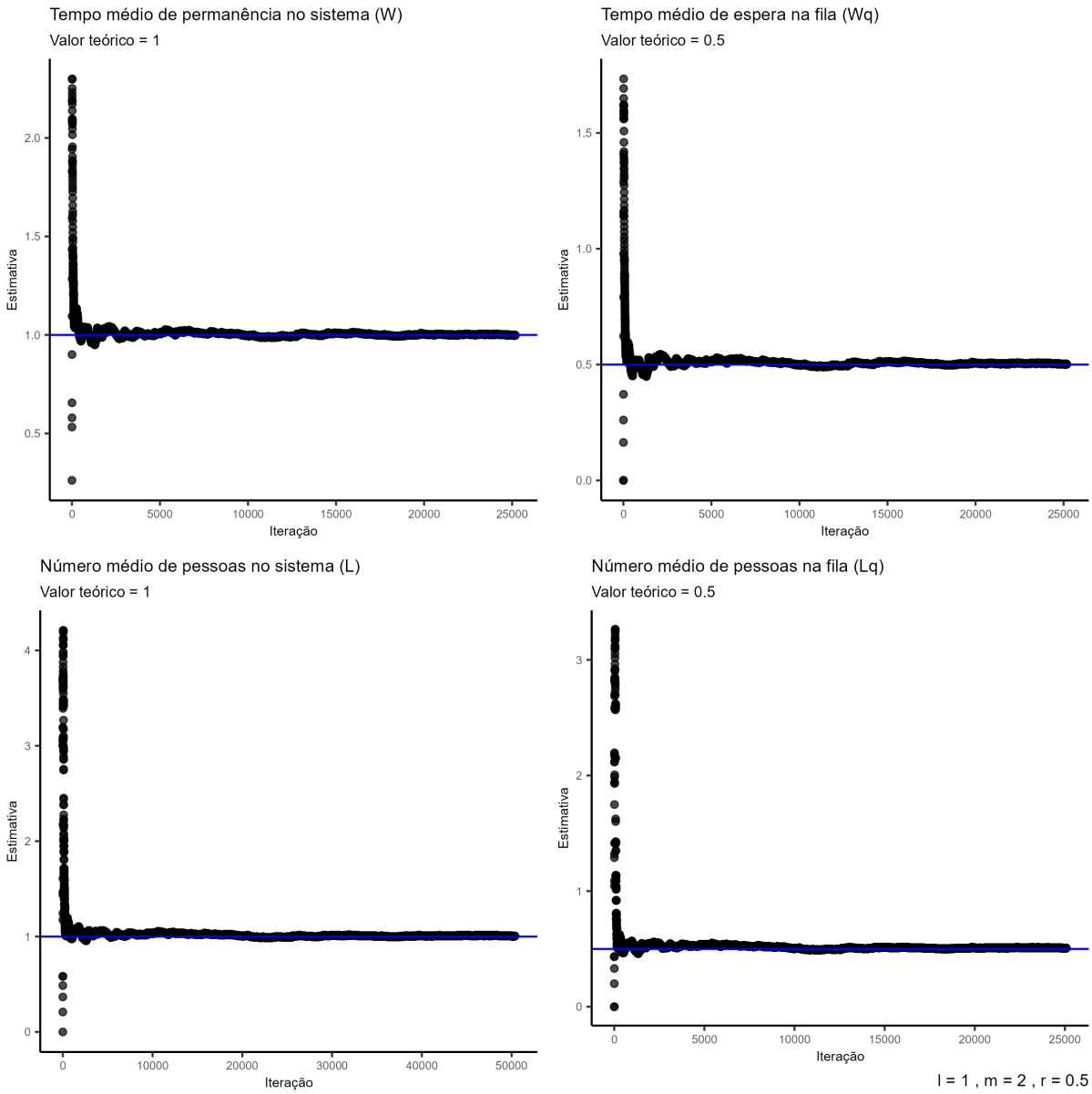

Figura 11: Sistema 4

APÊNDICE 4 – Gráficos M/M/1/k/*FIFO*

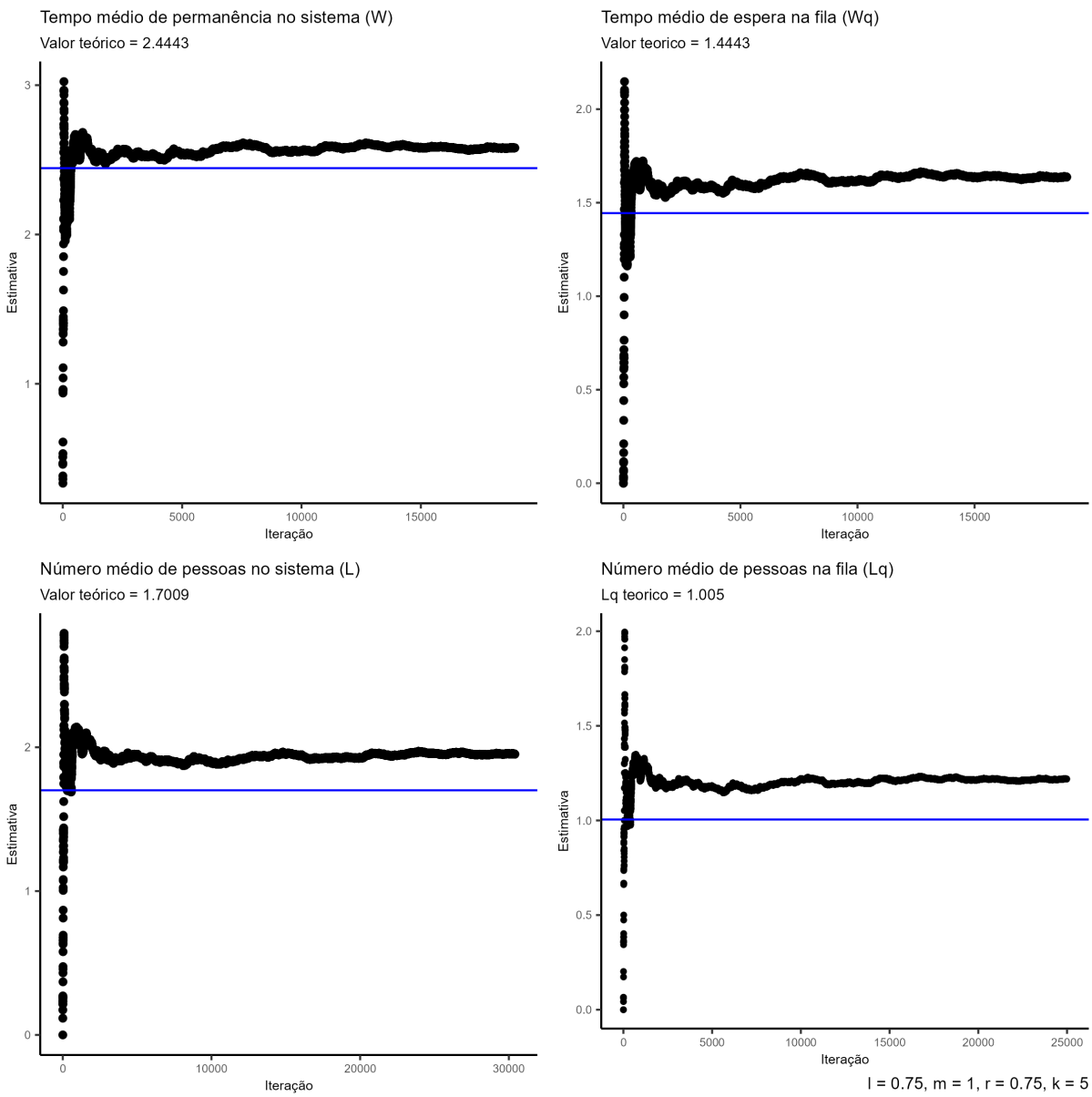

Figura 12: Sistema 1

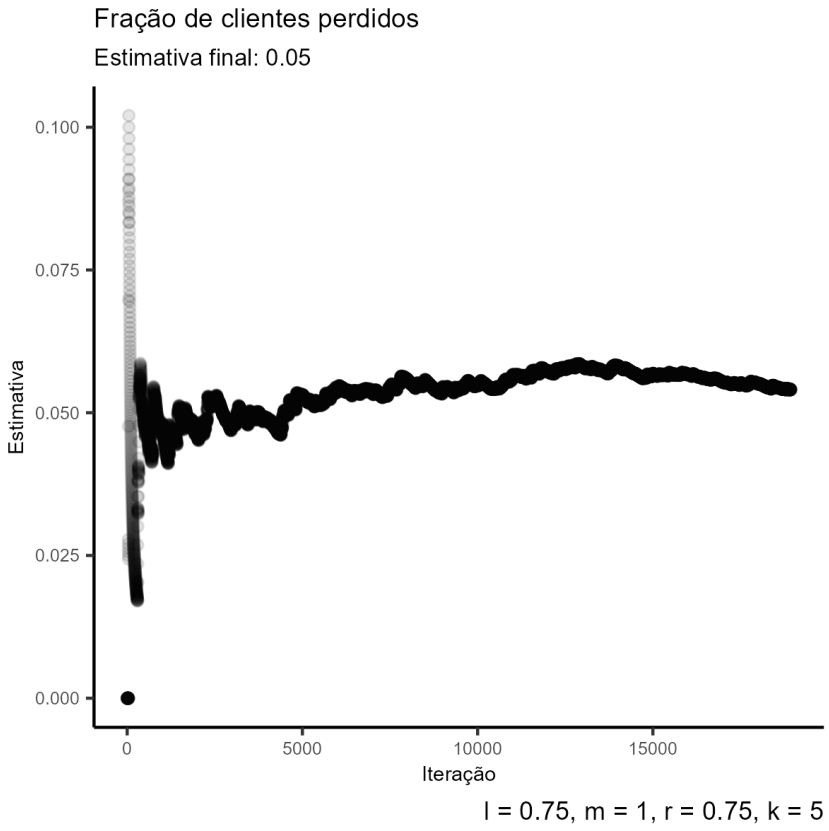

Figura 13: Sistema 1

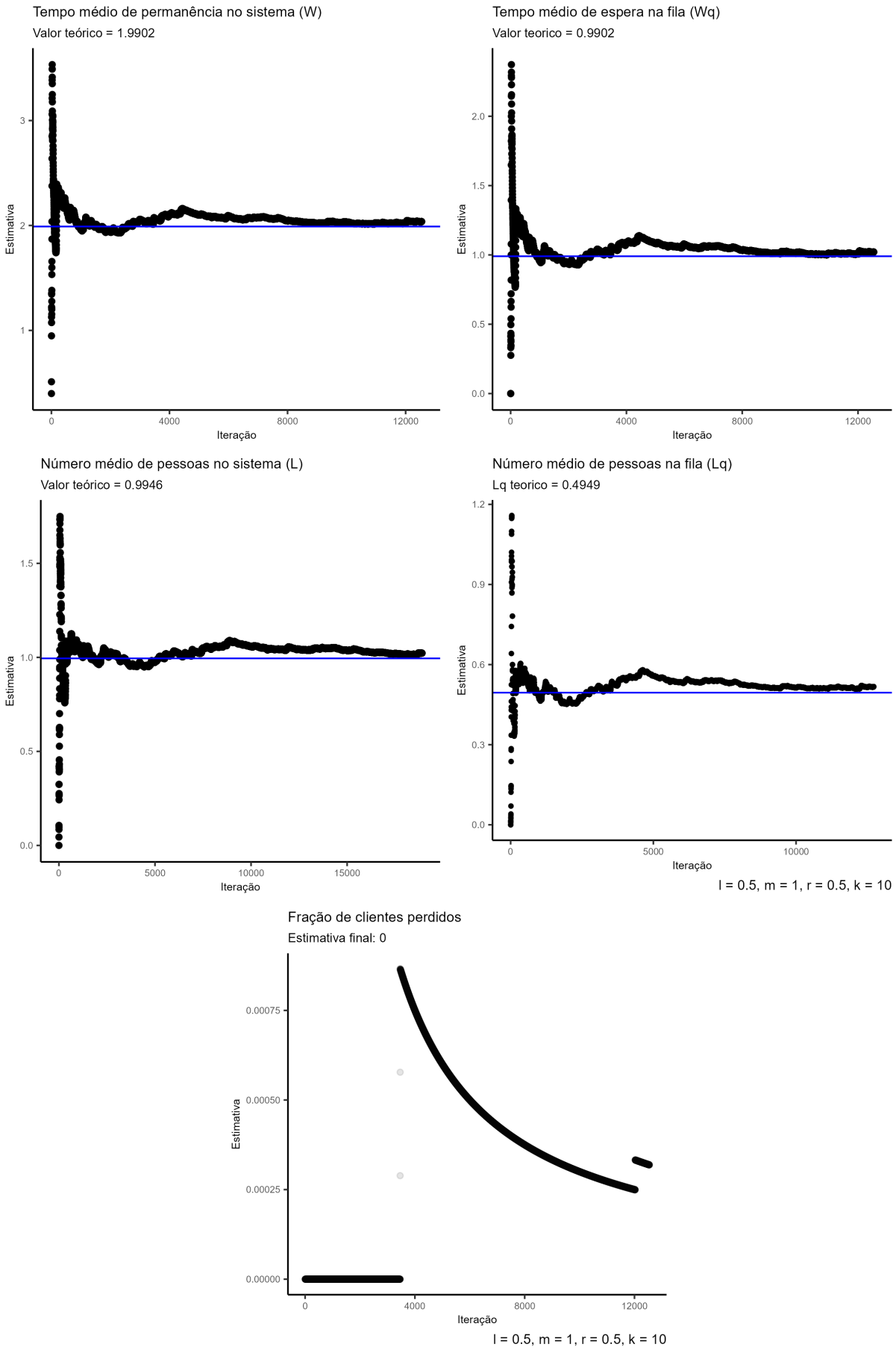

Figura 14: Sistema 2

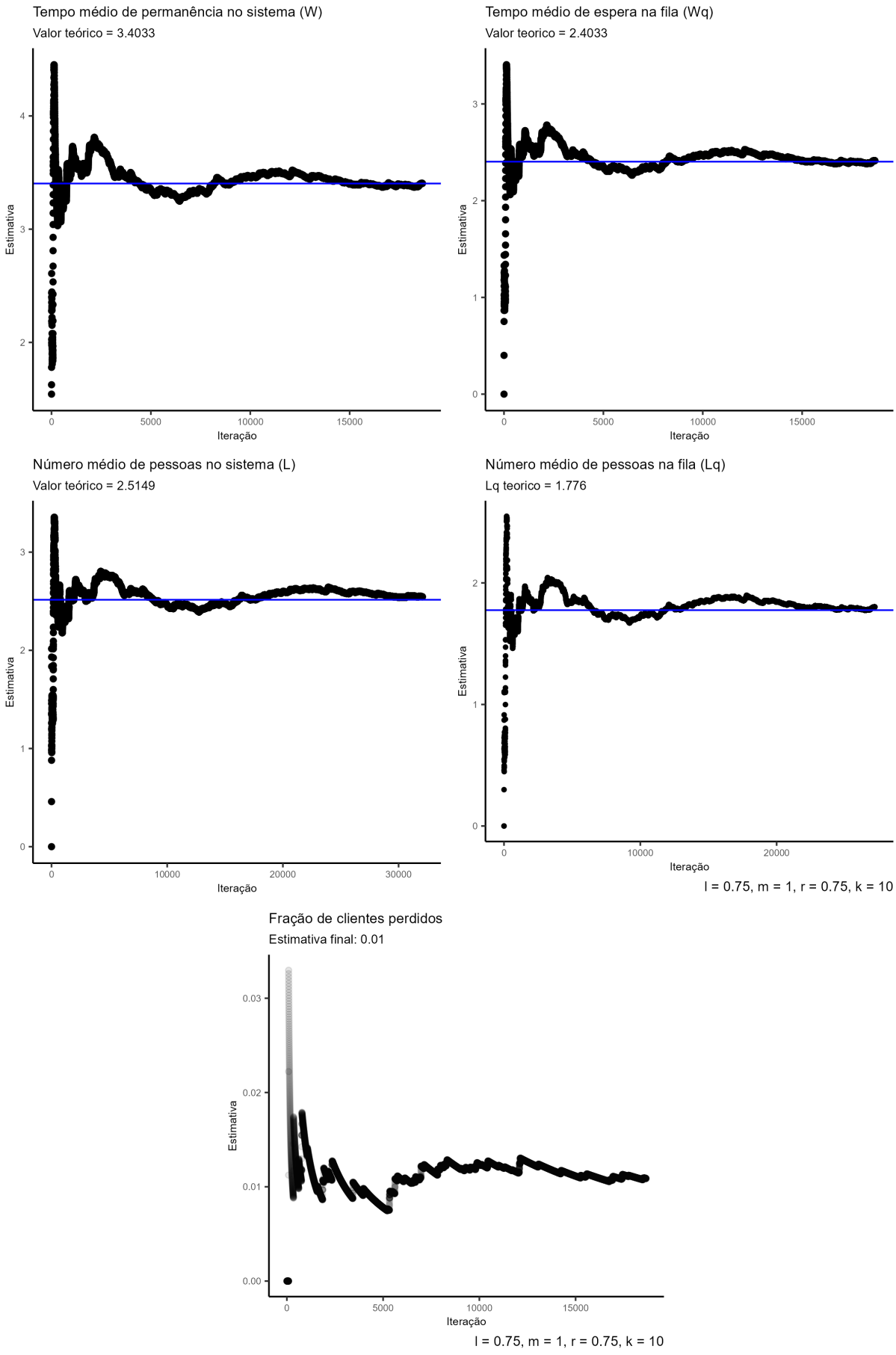

Figura 15: Sistema 3