Universidade de Brasília
Departamento de Estatística

LUCAS DE MORAES BASTOS

ESTUDO DE FILAS EM SERVIÇOS EMERGENCIAIS HOSPITALARES

Brasília
Outubro de 2020

LUCAS DE MORAES BASTOS

ESTUDO DE FILAS EM SERVIÇOS EMERGENCIAIS HOSPITALARES

Projeto apresentado para obtenção do título de Bacharel em Estatística

Universidade de Brasília
Departamento de Estatística

Orientador: Prof. Guilherme S. Rodrigues

Brasília
Outubro de 2020

Resumo

Este trabalho contém uma revisão de literatura de Teoria de Filas relativamente abrangente. Esse ramo de pesquisa possui variados modelos analíticos e modelos baseados em simulação computacional. O aspecto mais relevante foi a construção de um modelo computacional capaz de descrever o atendimento de emergência em um ambiente real, baseado no Hospital das Forças Armadas de Brasília. Dessa forma, foi desenvolvido um programa para análise de processos de espera, assim como um painel (*dashboard*) e uma animação do caminho percorrido por pacientes simulados.

Palavras-chave: Teoria de Filas. Simulação computacional. Processos de espera. Análise de atendimento médico-hospitalar.

Lista de Figuras

1 Histogramas dos tempos do sistema. 11
2 Família de distribuições Erlang com média igual a 3. 15
3 Histograma dos tempos entre chegada, com distribuição exata em vermelho. 19
4 Utilização e espera de um exemplo de atendimento em um hospital. 33
5 Fluxo de atendimento na emergência do HFA de Brasília (HFA Data & Care). 37
6 (a) Quantidade média de pacientes que dão entrada na emergência do HFA por dia e horário. (b) Tempo médio em cada fase de atendimento da emergência do HFA por dia e horário. 42
7 (a) Histograma do tempo de espera empírico para ser atendido na recepção. (b) Histograma do tempo de espera simulado para ser atendido na recepção. 44
8 (a) Desempenho do atendimento nos consultórios médicos de acordo com os diferentes cenários simulados, com 0 representando a estrutura original de atendimento e 1,2 e 3 representando as alterações estudadas na estrutura de atendimento. (b) Desempenho do atendimento na sala amarela de acordo com os diferentes cenários simulados, de maneira análoga à 8(a). 46
9 *Dashboard* de medidas de desempenho gerais da emergência do HFA. . . . 47
10 *Dashboard* de medidas de desempenho do atendimento médico de emergência do HFA. 48
11 Momento da animação que mostra o caminho percorrido pelos pacientes simulados de número 75 a 79. Observe que o paciente 77 já terminou seu atendimento enquanto os pacientes 75, 76 e 78 estão ainda sendo atendidos nos consultórios e o paciente 79 ainda está sendo atendido na recepção. . . 49

Lista de Tabelas

1 Medidas de desempenho para o modelo $M/M/c/K$ 10
2 Medidas de desempenho para exemplo de um sistema $M/M/1$. 12
3 Medidas de desempenho para exemplo de um sistema $G/M/1$. 20
4 Medidas de desempenho calculadas por simulação. 31
5 Seis observações de uma tabela referente ao atendimento na recepção gerada pelo modelo. 43
6 Seis observações de uma tabela referente ao atendimento nos consultórios médicos gerada pelo modelo. 43
7 *Summary* do tempo de espera empírico e do temo de espera simulado. . . . 43

8 Estrutura da força de trabalho do atendimento de emergência do hospital das forças armadas, contendo a estrutura original em vigência no funcionamento do hospital e três alterações simuladas, sendo D, o turno diurno e N, o noturno. 45
9 Medidas de desempenho do atendimento em cada estágio retornadas pelo modelo de acordo com a estrutura atual de atendimento de emergência do HFA de Brasília. 45

Sumário

1 Introdução . . . 5
2 Objetivos . . . 6
2.1 Objetivo Geral . . . 6
2.2 Objetivos Específicos . . . 6
3 Revisão de literatura . . . 7
3.1 Modelos básicos de Teoria de Filas . . . 7
3.1.1 $M/M/c/k$ . . . 8
3.1.2 Exemplo - $M/M/1$ . . . 10
3.2 Modelos avançados de Teoria de Filas . . . 12
3.2.1 Modelos com chegadas e/ou atendimento em lotes . . . 12
3.2.2 Modelos Erlangianos - $M/E_k/1$, $E_k/M/1$ . . . 14
3.2.3 Modelos Generalizados - $M/G/1$, $G/M/1$, $G/G/1$ . . . 16
3.2.4 Exemplo - $G/M/1$ . . . 19
3.3 Procedimentos de aproximação . . . 20
3.3.1 Limites e Inequações . . . 20
3.3.2 Aproximações . . . 21
3.4 Simulações . . . 23
3.4.1 Modelagem dos tempos do processo . . . 25
3.4.2 Análise dos resultados . . . 28
3.4.3 Validação do Modelo . . . 30
3.4.4 Exemplo - Simulação . . . 30
3.4.5 Exemplo - Sistema com múltiplos níveis de atendimento . . . 31
4 Metodologia . . . 34
4.1 Descrição probabilística do processo . . . 34
4.2 Materiais . . . 35
4.3 Algoritmo QDC . . . 37
4.4 Validação do programa implementado . . . 39
4.4.1 Comparação com *queuecomputer* . . . 39
5 Resultados . . . 41
5.1 Considerações sobre o banco de dados . . . 41
5.2 Análise Descritiva . . . 41
5.3 Funcionamento do Modelo Computacional . . . 42
5.4 Estudo para remanejamento mais eficiente da força de trabalho do hospital . . . 44
5.5 Aplicativo *Dashboard* . . . 46
5.6 Animação do caminho percorrido pelos pacientes . . . 48
6 Conclusão . . . 50
6.1 Sugestões para trabalhos futuros . . . 51

Referências 52
7 Apêndices 53
7.1 Código para simulação do problema da seção 3.4.5 - pacote *Simmer* 53
7.2 Descrição das variáveis do conjunto de dados do hospital HFA 56
8 Anexos 57
8.1 Demonstração - Função de distribuição acumulada do tempo de espera na fila ($W_q(t)$) para o modelo $M/M/c/k$ 57
8.2 Outros modelos básicos 58
8.3 Demonstrações - modelo $M^{[X]}/M/1$ 67
8.4 Demonstrações - $M/G/1$ 69
8.5 Demonstrações - $G/M/1$. 70

1 Introdução

Filas são um aspecto onipresente e de acentuada relevância na vida em sociedade. Há filas de carros no trânsito, filas em bancos, filas no serviço público, filas em UTI, as quais viraram notícia nestes tempos de pandemia, para citar apenas algumas. Diversas atividades, intrinsecamente, envolvem um ou vários processos de espera (filas), o que é o caso no atendimento de emergência de um hospital, foco deste trabalho.

A partir da análise de dados (como tempos de chegada dos pacientes, tempos de atendimento, número de enfermeiros na triagem e médicos nos consultórios ao longo do tempo), podemos construir modelos probabilísticos da Teoria de Filas e estudos de simulação que descrevem o comportamento do processo. Tais modelos permitem identificar gargalos e ineficiências do processo, provendo recomendações de gestão para diminuir o tempo de espera e otimizar a alocação dos recursos. Isso leva, em última instância, à diminuição de custos operacionais e na melhora dos serviços prestados e na satisfação do usuário, evitando, ainda, a perda de clientes que desistem de esperar. É possível, por exemplo, mensurar, antes mesmo da implementação, o efeito esperado da adição de um posto de triagem no tempo médio de espera do paciente (considerando-se a infraestrutura em operação), de modo a avaliar se esta seria uma intervenção efetiva ou mero desperdício de recursos. O monitoramento contínuo das filas, com base em medidas de desempenho (tempo médio de espera, taxa de ocupação, número médio de pessoas na fila, entre outros), também pode ser visto como parte central do controle de qualidade dos serviços prestados.

Tornar filas mais rápidas e eficientes está em concordância com a dignidade da pessoa humana, fundamento expresso da Constituição Federal de 1988. Além disso, a demora excessiva no atendimento de emergência de hospital é falha de serviço tipificada no artigo 14 do Código de Defesa do Consumidor (Lei 8.078/1990). Alguns municípios vêm aprovando leis que estabelecem tempo máximo de espera para atendimento em emergência de hospitais públicos e particulares, como a lei municipal Nº 6.358 de Maceió (AL) e a lei Nº 11405 do município de Sorocaba, SP, que "dispõe sobre o tempo máximo de espera em prontos socorros que atendem pacientes conveniados". Há casos que chegaram ao STJ como dano moral coletivo em casos de espera em agências bancárias.

Este trabalho também se identifica intimamente com o objetivo essencial da pesquisa científica: oferecer soluções a questões que causam transtornos à sociedade. Frequentemente, há um distanciamento entre a academia, a sociedade e o mercado. A integração desses setores é benéfica a todas as partes: a Universidade gera pesquisa que contribui diretamente para a comunidade; os usuários do produto da pesquisa desfrutam de maior bem-estar e as empresas ou instituições reduzem custos e alavancam seus negócios. O

aluno pesquisador também se beneficia ao experimentar a atividade profissional de maneira mais autêntica e aprofundar seus conhecimentos técnicos.

2 Objetivos

2.1 Objetivo Geral

Subsidiar a equipe gestora com vistas a otimizar a alocação dos recursos aplicados e aumentar a qualidade do atendimento em todas as suas fases: triagem, identificação do paciente (recepção), consulta, exame, etc. Para este intento, trabalhamos em duas frentes, a saber, construção e calibração de um modelo computacional capaz de responder a uma série de questões relativas ao sistema e criação de ferramentas de visualização do processo, facilitando a compreensão do fluxo de pacientes e a identificação de possíveis gargalos.

2.2 Objetivos Específicos

- Identificar pontos do processo (recepção, triagem, consulta, etc.) com maior impacto relativo no tempo de espera e no tamanho das filas;
- Propor critérios para o dimensionamento da força de trabalho (médicos, recepcionistas, enfermeiros) ao longo do dia;
- Implementar e disponibilizar um sistema de monitoramento contínuo e visualização do desempenho do sistema;
- Construir o modelo que descreve a formação das filas.

3 Revisão de literatura

3.1 Modelos básicos de Teoria de Filas

Os modelos que descrevem processos de filas resultam da aplicação de resultados de processo de nascimento e morte, demonstrado como sendo um tipo de cadeia de Markov de parâmetros contínuos (Gross e Harris, 1998).

A notação utilizada foi proposta por Kendall (1953) sendo a mais utilizada hodiernamente. Ela é escrita pela expressão A/B/C/D/E, em que "A" e "B" são as distribuições dos tempos entre chegadas sucessivas e dos tempos de atendimento ou de serviço, respectivamente, que devem ser independentes; "C" é o número de postos de atendimento em paralelo; "D" é a capacidade física do sistema; por fim, "E" se refere a disciplina de filas utilizada. Para "A" e "B" as notações mais utilizadas são: "M", que é exponencial (*memoryless*), "D", determinístico e "G", geral. Para "C" e "D", os valores podem ser de 1 a ∞; para "E", há diversas formas de disciplinas de filas, sendo a mais comum a FIFO (*First in first out*); D e E são indicados também pela sua ausência, ou seja, M/M/1, é um modelo com chegadas e atendimentos exponenciais, único canal de serviço, capacidade do sistema infinita e disciplina de atendimento FIFO.

É importante destacar que a suposição de independência entre os tempos de chegada e os tempos de atendimento é facilmente violada. Se ocorre uma queda na rede de energia do estabelecimento, por exemplo, todos os usuários deverão esperar, acumulando a quantidade de pessoas em espera e atrasando o serviço. Outra situação que descaracterizaria a independência seria a ocorrência de um grande número de chegadas de usuários de uma só vez, por exemplo, se ocorresse um acidente envolvendo várias vítimas e muitas delas fossem para o mesmo hospital.

Mesmo os modelos analíticos mais avançados que serão revisados possuem essa suposição limitante de independência entre os tempos, constituindo-se uma das grandes razões a favor do uso de modelos de simulação, capazes de modelar situações reais complexas.

O primeiro modelo, analisado na seção 3.1.1, é o $M/M/c/K$, ou seja, com distribuição Poisson de tempos de chegada, tempos de serviço Exponencial, c canais de atendimento e limitação de K clientes no sistema. A partir desse modelo, mais geral, podem-se especificar algumas condições para se chegar aos outros modelos básicos.

Fórmulas de Little

Uma das relações mais poderosas e de extrema utilidade em Teoria de Filas foi

desenvolvida por John D. C. Little na década de 1960 (Little, 1961). Ele relacionou os tamanhos médios do sistema aos tempos médios de espera, ambos no estágio estacionário, isto é, quando as distribuições de probabilidade que descrevem os tempos do modelo (entre chegada e de serviço) são invariantes à passagem do tempo.

Assim, o número médio de usuários em um sistema é igual ao produto da taxa média de ingresso pelo tempo médio de permanência de um usuário no mesmo, analiticamente

$$L = E(\Delta)W,$$

em que $E(\Delta)$ é a taxa média de ingressos no sistema ou o parâmetro da distribuição dos tempos de chegada.

Outras fórmulas de Little, que expressam a relação entre as medidas de desempenho do sistema são:

$$W = W_q + E(A),$$

$$W_q = \frac{L_q}{E(\Delta)} \quad \text{e}$$

$$L_q = L - E(\Delta)E(A),$$

em que A é o tempo que um usuário permanece em atendimento; W é o tempo médio de espera no sistema, incluindo o tempo na fila e o tempo em atendimento; W_q é o tempo médio de espera na fila; e L_q, o tamanho médio da fila.

Essas fórmulas são muito úteis por que o conhecimento de uma medida de desempenho implica no conhecimento das outras, desde que saibamos as taxas de chegada e de atendimento. Por exemplo, para o modelo mais simples, *M/M/1*, tem-se $E(\Delta) = \lambda$ e $E(A) = \frac{1}{\mu}$, em que λ e μ são a taxa de chegada dos usuários ao sistema e a taxa do tempo de atendimento, respectivamente.

3.1.1 $M/M/c/k$

O modelo $M/M/c/k$ é definido pelas seguintes suposições:

- tempos entre chegadas sucessivos e tempos de atendimento distribuídos exponencialmente;
- existência de c postos de atendimento;
- fila de espera com limite K de ocupação;

- ordem de atendimento seguindo a ordem de chegada dos usuários ao sistema (FIFO).

O processo de chegada é Poisson, implicando que os tempos entre chegada são exponenciais (Gross e Harris, 1998). Os tempos de serviço também são exponenciais, logo, tem-se um processo de nascimento e morte. Portanto, as taxas de chegada e de atendimento são constantes, dadas, respectivamente, por

$$\lambda_n = \begin{cases} \lambda, & \text{se } 0 \leqslant n < k \\ 0, & \text{se } n \geqslant k. \end{cases}$$

e

$$\mu_n = \begin{cases} n\mu, \text{ se } 1 \leqslant n < c \\ c\mu, \text{ se } c \leqslant n \leqslant k. \end{cases}$$

Em qualquer processo Markoviano que se encontra em estágio estacionário, a probabilidade do sistema estar no estado n no instante t (ou seja, o número de usuários ser n no instante t) é

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

Pela distribuição estacionária de processos de nascimento e morte (Gross e Harris, 1998), tem-se

$$P_n = \begin{cases} \frac{\lambda^n}{n!\mu^n}\mathrm{P}_0, \text{ se } 1 \leqslant n < c \\ \frac{\lambda^n}{c^{n-c}c!\mu^n}\mathrm{P}_0, \text{ se } c \leqslant n \leqslant k. \end{cases}$$

e

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{K} \frac{\lambda^n}{c^{n-c}c!\mu^n} \right)^{-1}, \ \forall \ n \geq 1.$$

Para maiores detalhes sobre o desenvolvimento das fórmulas, ver demonstrações contidas no Anexo, Seções 8.1 e 8.2.

Medida	Fórmula
L_q	$L_q = \frac{P_0 r^c \rho}{c!(1-\rho)^2}\left[1-\rho^{K-c+1}-(1-\rho)(K-c+1)\rho^{K-c}\right]$
L	$L_q + r(1-p_k)$
W	$\frac{L}{\lambda(1-p_k)}$
W_q	$\frac{L_q}{\lambda_{eff}}$
$W_q(t)$	$1-\sum_{n=c}^{K-1} q_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}$
$P(T_q > t)$	$1 - W_q(t)$

Tabela 1: Medidas de desempenho para o modelo $M/M/c/K$

L_q: número médio de usuários na fila; L: número médio de usuários no sistema; W: tempo médio de permanência no sistema; W_q: tempo médio de espera na fila; $W_q(t)$: função de distribuição acumulada do tempo de espera na fila; $P(T_q > t)$: probabilidade do tempo de espera na fila ser maior do que um tempo $t > 0$.

Como há limitação na capacidade do espaço de espera, as fórmulas de Little sofrem algumas modificações (Gross e Harris, 1998). Uma fração p_k das chegadas não entra no sistema, porque chegaram quando já não havia mais espaço. Assim, a taxa de chegada efetiva é ajustada de acordo com esse fator limitante, senda dada, agora, por $\lambda_{eff} = \lambda(1-p_k)$. Assim, a relação entre L e L_q é reajustada. A quantidade $r(1-p_k)$ deve ser menor que c, tendo em vista que o número médio de clientes em atendimento deve ser menor do que o número total de servidores disponíveis, o que indica a definição de $\rho_{eff} = \lambda_{eff}/c\mu$, que deve ser menor que 1, apesar de não existir tal restrição para $\rho = \lambda/c\mu$. Além disso, $r = \lambda/\mu$ e $\rho = r/c$.

Casos particulares

Os outros modelos básicos, com as distribuições dos tempos entre chegada e de atendimento exponenciais, podem ser deduzidos a partir do modelo $M/M/c/K$, variando-se os valores de c e de K. O modelo mais simples, por exemplo, sem limitações na capacidade do sistema e na quantidade de atendentes, $M/M/1$, é um $M/M/c/K$ com $c \to \infty$ e $K \to \infty$

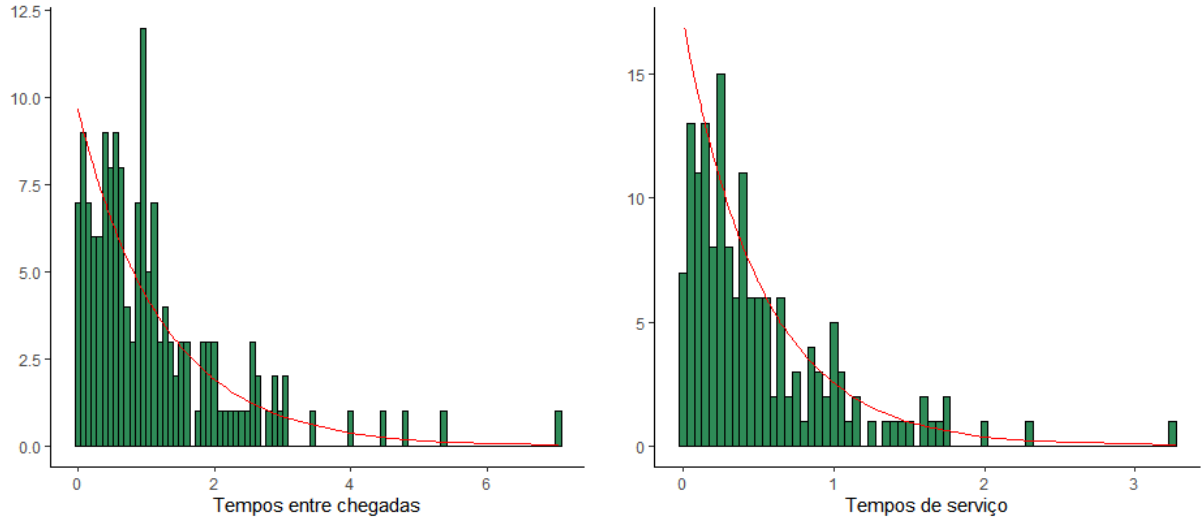
3.1.2 Exemplo - $M/M/1$

Suponha que determinado hospital oftalmológico possui apenas uma recepcionista, que atende aos pacientes de acordo com a ordem de chegada.

A fim de explorar as propriedades dos diversos modelos considerados neste trabalho, foram gerados dados por simulação de um modelo especificado da seguinte forma: distribuição exponencial dos tempos de atendimento, com taxa $\mu = 2$ e distribuição Gama dos

tempos entre chegadas, com parâmetro de forma $\alpha = 1.23$ e taxa $\beta = 1$. Observe que os parâmetros utilizados, assim como as distribuições de probabilidade, são desconhecidos em estudos aplicados. O objetivo, portanto, está em, a partir dos dados observados (simulados, neste caso), construir um modelo que melhor aproxime o processo ("desconhecido") que os gerou.

Foram obtidos os tempos de chegada e de serviço de $n = 158$ pacientes. Dois testes de Kolmogorov-Smirnov foram realizados para testar as hipóteses de que os tempos entre chegadas e de que os tempos de serviço são exponenciais, obtendo-se os p-valores respectivos de 0.1055 e 0.9548. Logo, não se rejeita que os tempos sejam distribuídos exponencialmente, a 5% de significância. Ainda foram comparados os histogramas com as distribuições exatas



(a) (b)

Figura 1: Histogramas dos tempos do sistema.

As estimativas obtidas por *bootstrap* dos dados gerados resultaram nos valores $\lambda = 1.119$ e $\mu = 1.951$, obedecendo à restrição $\rho = \lambda/\mu = 0.573 < 1$, que corresponde à intensidade do tráfico no sistema.

A probabilidade de haver $n = 0$ chegadas é dada por $P_0 = 1 - \rho = 0.541$. A partir disso, podem ser obtidas as medidas de desempenho do sistema, apresentadas na Tabela 2.

Observa-se que as estimativas aproximam-se dos valores reais, mas os intervalos de confiança são grandes. Veremos que a estimação ainda pode ser melhor, quando retomarmos este exemplo nas Seções 3.2.4 e 3.4.4. Isto porque o modelo ajustado está mal especificado.

Medida	Fórmula	Valor real	Valor estimado	I.C. (95%)
L	$\frac{\rho}{1-\rho}$	1	1.343	[0.698, 3.205]
L_q	$\frac{\rho^2}{(1-\rho)}$	0.5	0.77	[0.287, 2.443]
W	$\frac{1}{(\mu-\lambda)}$	1 min	1.2 min	[0.736, 2.489]
W_q	$\frac{\rho}{(\mu-\lambda)}$	0.5 min	0.688 min	[0.302, 1.897]
$P(N \geq 3)$	ρ^k	0.125	0.188	[0.069, 0.443]
$P(T_q > 2)$	$\rho e^{-(\mu-\lambda)t}$	0.068	0.108	[0.027, 0.341]

Tabela 2: Medidas de desempenho para exemplo de um sistema $M/M/1$.

3.2 Modelos avançados de Teoria de Filas

Gradualmente, foram desenvolvidos modelos que relaxam as suposições necessárias para a utilização dos métodos analíticos, com a introdução do número de chegadas como variáveis aleatórias e, depois, considerando-se outras distribuições que definam os tempos de chegada e os tempos de serviço.

3.2.1 Modelos com chegadas e/ou atendimento em lotes

$M^{[X]}/M/1$

Com este modelo, há a possibilidade de haver mais de uma chegada de uma vez, descaracterizando-se o processo de nascimento e morte, mas ainda sendo um processo Markoviano. Ou seja, além da suposição de que as chegadas formam um processo de Poisson (tempos entre chegadas, consequentemente, distribuídos exponencialmente) é assumido que a quantidade de clientes que chegam em qualquer momento é uma variável aleatória X, que assume qualquer valor inteiro positivo com probabilidade c_x. Note-se que este modelo ainda é Markoviano no sentido de que o comportamento futuro é uma função apenas do presente e não do passado (Gross e Harris, 1998).

Se λ_x é a taxa de chegada de um processo de Poisson com "lotes de chegada" de tamanho X, então $c_x = \lambda_x/\lambda$, em que λ é a taxa de chegada composta de todos os lotes, sendo $\lambda = \sum_{i=1}^{\infty} \lambda_i$.

Este processo total, que surge pela sobreposição do conjunto de processos de Poisson com taxas $\{\lambda_x,\, x = 1,\, 2,\, \dots\}$, é um processo de Poisson múltiplo ou composto.

As probabilidades de n chegadas ao sistema são dadas por

$$P_0 = 1 - \rho$$
$$P_n = (1-\rho)[\alpha + (1-\alpha)\rho]^{n-1}[(1-\alpha)\rho] \;\; (n > 0),$$

em que $c_x = (1-\alpha)\alpha^{x-1}$, $(0 < \alpha < 1)$, são os tamanhos dos lotes de chegada distribuídos geometricamente.

Os cálculos e demonstrações para este modelo também se encontram no Anexo, na Seção 8.3. O número médio de usuários no sistema é dado por:

$$L = \frac{r\{E[X] + E[X^2]\}}{2(1-\rho)} = \frac{\rho + rE[X^2]}{2(1-\rho)}.$$

$\rho < 1$ é condição necessária e suficiente para a estacionariedade.

As outras medidas de eficiência do sistema são encontradas facilmente utilizando o resultado $L_q = L - \rho$ e as fórmulas de Little (Gross e Harris, 1998).

Além disso, podem-se estender esses resultados para o modelo $M^{[X]}/M/c$, de maneira similar ao que foi feito para os modelos $M/M/1$ e $M/M/c$.

$M/M^{[K]}/1$

Para este modelo, considera-se que K clientes são servidos simultaneamente, formando "lotes de atendimento", sendo o tempo para cada lote distribuído exponencialmente. Portanto, também não se caracteriza como processo de nascimento e morte, mas é Markoviano.

As equações balanceadas estocásticas são (Gross e Harris, 1998),

$$\begin{aligned} 0 &= -(\lambda+\mu)P_n + \mu P_{n+K} + \lambda P_{n-1} \qquad (n \geq 1), \\ 0 &= -\lambda P_0 + \mu P_1 + \mu P_2 + \cdots + \mu P_{K-1} + \mu P_K. \end{aligned}$$

A primeira equação pode ser reescrita utilizando-se a notação operador comum,

$$[\mu D^{K+1} - (\lambda+\mu)D + \lambda]P_n = 0.$$

Dessa forma, se $(r_1, \ldots, r_{K+1})$ são as raízes do operador ou equação característica, então

$$P_n = \sum_{i=1}^{K+1} C_i r_i^n \qquad (n \geq 0).$$

Desde que $\sum_{n=0}^{\infty} P_n = 1$, cada r_i deve ser menor que um ou $C_i = 0$ para todo r_i que não seja menor que 1. De fato, pode ser verificado, utilizando-se o Teorema de Rouché, que há somente uma raiz, r_0, no intervalo $(0, 1)$. Então

$$P_n = Cr_0^n \qquad (n \geq 0,\ 0 < r_0 < 1).$$

Utilizando a condição limitante de que $\sum P_n$ deve totalizar um, tem-se que $C =$

$P_0 = 1 - r_0$. Logo,

$$P_n = (1 - r_0)r_0^n.$$

Como a solução estacionária possui a mesma forma geométrica que $M/M/1$, com r_0 no lugar de ρ, podem-se obter imediatamente as medidas de eficiência:

$$L = \frac{r_0}{1 - r_0}, \qquad L_q = L - \frac{\lambda}{\mu}$$

e

$$W = \frac{r_0}{\lambda(1 - r_0)}, \qquad W_q = W - \frac{1}{\mu}.$$

3.2.2 Modelos Erlangianos - $M/E_k/1$, $E_k/M/1$

Até agora, todos os modelos probabilísticos estudados assumem distribuição de chegadas Poisson (tempos entre chegadas exponenciais) e tempos de serviço/atendimento exponenciais e variações que consistem na atribuição da distribuição exponencial para conjuntos ("lotes") de clientes que chegam ou são atendidos simultaneamente. Mas, em muitas situações práticas, a suposição de exponencialidade é muito limitante.

A Distribuição de Erlang

Considere uma variável aleatória T que possua densidade Gama. Uma classe especial dessas distribuições, em que $\alpha = k$ e $\beta = 1/k\mu$, sendo k um inteiro positivo arbitrário e μ uma constante positiva qualquer, é a família Erlang de distribuições de probabilidade:

$$f(t) = \frac{(\mu k)^k}{(k-1)!} t^{k-1} e^{-k\mu t} \quad (0 < t < \infty).$$

A família de distribuições de probabilidade de Erlang possui maior flexibilidade que a exponencial, que é, de fato, um caso particular da Erlang para $k = 1$.

Apesar de ser não-Markoviana (dentre as distribuições contínuas apenas a exponencial possui a propriedade de Markov), a distribuição erlangiana possui relação próxima com a distribuição exponencial: observe que a soma de k variáveis aleatórias exponenciais IID com média $1/k\mu$ resulta em uma Erlang tipo k. Tal relação permite a análise de modelos de filas com tempos de chegada e/ou de serviço distribuídos de acordo com a Erlang.

Processos muito bem descritos pela distribuição de Erlang são os que possuem várias fases de atendimento com tempos exponenciais e mesma taxa média. Além dessa limitação, há outra relacionada ao ingresso de novos usuários: um novo usuário é atendido

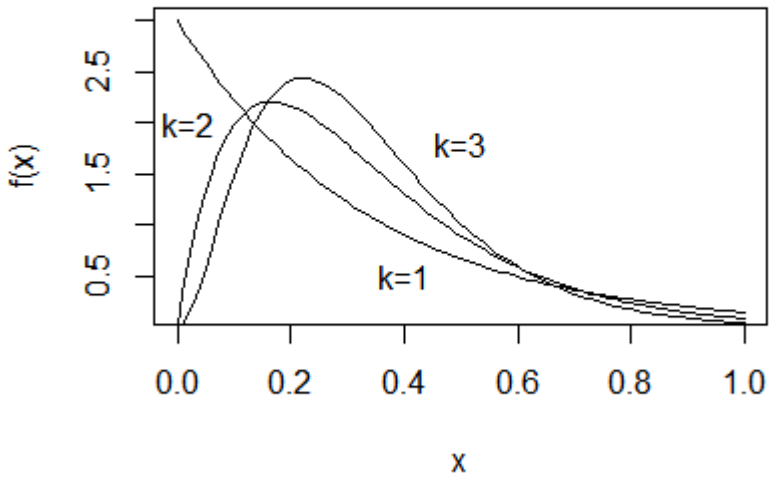


Figura 2: Família de distribuições Erlang com média igual a 3.

apenas quando todas as fases de atendimento são completadas para o usuário anterior (Gross e Harris, 1998).

Modelo com serviço Erlang ($M/E_k/1$)

Neste modelo, os tempos de atendimento têm distribuição de Erlang tipo k, e pode-se fazer uma equivalência com o modelo $M/M^{[K]}/1$, em que cada lote de chegada traz $K = k$ fases e a taxa de serviço é substituída por $k\mu$. Isso leva aos resultados (Gross e Harris, 1998):

$$W_q = \frac{k+1}{2k}\frac{\rho}{\mu(1-\rho)} \qquad (\rho = \lambda/\mu)$$

$$L_q = \lambda W_q = \frac{k+1}{2k}\frac{\rho^2}{1-\rho},$$

lembrando que $L = L_q + \rho$ e $W = L/\lambda = W_q + 1/\mu$.

As probabilidades estácionárias são: $P_0 = 1-\rho$, pois há apenas um canal de serviço, sendo o atendimento de apenas um usuário por vez, e $P_n = \sum_{j=(n-1)k+1}^{nk} p_j^{(P)}$ $(n \geq 1)$, em que $p_n^{(P)}$ representa a probabilidade de n usuários em cada lote de chegada.

Modelo com chegada Erlang ($E_k/M/1$)

Da mesma forma que no modelo anterior, podem-se utilizar os resultados do modelo $M^{[K]}/M/1$, considerando-se uma chegada que passa por k fases, cada uma com um tempo médio de $1/k\lambda$. Ou seja, assume-se que os tempos entre chegada possuem distribuição de

Erlang tipo k com média $1/\lambda$.

A probabilidade no estado estacionário é dada pelo somatório das probabilidades de haver n fases de chegada ao sistema no estágio estacionário, $P_n = \sum_{j=nk}^{nk+k-1} p_j^{(P)}$, em que $p_j^{(P)} = \rho(1 - r_0)r_0^{j-k}$, para $j \geq k$, e, portanto,

$$P_n = \rho(1 - r_0^k)(r_0^k)^{n-1}.$$

Vê-se que essa é uma distribuição geométrica, como em $M/M/1$, mas com r_0^k como o multiplicador ao invés de ρ. Logo,

$$L = \rho(1 - r_0^k) \sum_{n=1}^{\infty} n(r_0^k)^{n-1} = \frac{\rho(1 - r_0^k)}{(1 - r_0^k)^2} = \frac{\rho}{1 - r_0^k}.$$

As outras medidas são facilmente obtidas pelas fórmulas de Little, e a função de distribuição acumulada do tempo de permanência na fila é obtida de maneira similar ao que é feito nos modelos básicos, e é dada por

$$W_q(t) = 1 - r_0^k e^{-\mu(1-r_0^k)t} \ (t \geq 0).$$

O modelo $E_j/E_k/1$ é um modelo mais geral, da forma $GE_j/GE_k/1$. Não foi incluído aqui porque o modelo mais geral é descrito $G/G/1$ é descrito na Seção 3.2.3.

3.2.3 Modelos Generalizados - $M/G/1$, $G/M/1$, $G/G/1$

Para os modelos mais gerais, quaisquer variáveis aleatórias podem ser escolhidas para representar os tempos do processo.

M/G/1

Neste modelo, tem-se tempo entre chegadas exponencial e tempos de serviço descritos por variáveis aleatórias IID com distribuição de probabilidade arbitrária. Pode-se demonstrar que é um processo Markoviano (Gross e Harris, 1998).

Ao calcular os valores esperados da variável aleatória X_n que representa a quantidade no sistema no n-ésimo ponto de saída (ver Anexo, Seção 8.4), pode-se obter o resultado

$$L = \rho + \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1 - \rho)}.$$

Pelo uso das fórmulas de Little podem ser encontradas as outras medidas. Portanto, há apenas que saber λ, a média e a variância das distribuições de chegada e de serviço.

As probabilidades do estado estacionário neste modelo são encontradas a partir das probabilidades π_n de haver n no sistema em determinado tempo de saída, que, neste caso, são iguais a P_n (Gross e Harris, 1998). Ver Anexo, Seção 8.4.

$$\pi_i = \pi_0 k_i + \sum_{j=1}^{i+1} \pi_j k_{i-j+1} \qquad (i = 0, 1, 2, \dots)$$
$$\pi_0 = 1 - \rho$$

Se existirem mais canais de serviço, ou seja, o caso do modelo $M/G/c$, há um resultado geral que relaciona o k-ésimo momento fatorial do tamanho do sistema e o k-ésimo momento regular da espera no sistema, dada por

$$L_{(k)} = \lambda^k W_k$$

G/M/1

Agora, pressupõe-se que os tempos de serviço são exponenciais com taxa média μ e assume-se apenas que os tempos de chegada são IID de acordo com alguma distribuição de probabilidade qulaquer. Uma abordagem de cadeia de Markov embutida também é utilizada aqui para se obterem os resultados e, de maneira similar ao modelo anterior, obtém-se as equações estacionárias usuais (Gross e Harris, 1998).

Os próximos resultados são análogos ao modelo básico $M/M/1$ com a diferença de que os tamanhos médios são válidos apenas para os momentos de chegada, indicado por $L^{(A)}$.

$$L^{(A)} = \frac{r_0}{1 - r_0} \qquad e \qquad L_q^{(A)} = \frac{r_0^2}{1 - r_0}.$$

em que r_0 é a única raiz que soluciona $z = A^*[\mu(1 - z)]$, sendo $A(z)$ a transformada Laplace-Stieltjes (LST) da f.d.a. dos tempos entre chegada.

As funções do tempo de espera na fila e do total de espera no sistema são dadas por

$$W_q(t) = 1 - r_0 e^{-\mu(1-r_0)t} \qquad (t \geq 0),$$
$$W(t) = 1 - e^{-\mu(1-r_0)t} \qquad (t \geq 0),$$

com valores médios

$$W_q = \frac{r_0}{\mu(1 - r_0)} \qquad e \qquad W = \frac{1}{\mu(1 - r_0)}.$$

Para o caso com múltiplos canais de serviço, $G/M/c$, reescreve-se

$$W_q(t) = 1 - \frac{q_c}{1 - r_0} e^{-c\mu(1-r_0)t}$$

com média

$$W_q = \frac{q_c}{c\mu(1 - r_0)^2}$$

G/G/1

Apesar de ser muito vago em sua estrutura, pode-se chegar a um resultado geral (Gross e Harris, 1998) para filas com chegadas e atendimentos distribuídos arbitrariamente, derivado de uma equação integral do tipo Wiener-Hopf para a distribuição estacionária do tempo de espera na fila de um cliente qualquer. Essa equação é, em maior parte, devida a Lindley (1958) e, por isso, possui seu nome.

Equação de Lindley:

$$W_q(t) = -\int_0^\infty W_q(y) dU(t - y) \qquad (0 \geq t < \infty),$$

em que $U(x)$ é a convolução de S e $-T$:

$$U(x) = \int_{max(0,x)}^\infty B(y) dA(y - x),$$

e $A(x)$ e $B(x)$ são as funções de ditribuição acumulada das v.a.'s do tempo de entre chegadas e dos tempos de serviço, T e S, respectivamente.

Resolvendo a equação de Lindley, obtém-se a transformada de Laplace para a função do tempo de espera na fila, denotada por $\bar{W}_q(s)$:

$$\overline{W}_q(s) = \frac{\overline{W}_{\overline{q}}(s)}{A^*(-s)B^*(-s) - 1}$$

em que $W_{\overline{q}}(s)$ é a parte da f.d.a. associada aos valores negativos de $W_q^{(n)} + s - T$ quando há tempo ocioso entre o n-ésimo e o cliente $(n + 1)$, dada por

$$W_{\overline{q}}(t) = \int_{-\infty}^t W_q(t - x) dU(x) \qquad (t < 0)$$

Portanto, dado qualquer par de f.d.a.'s dos tempos entre chegada e dos tempos de atendimento, $\{A(t),\ B(t)\}$, em um sistema $G/G/1$, pode-se, teoricamente, encontrar a transformada de Laplace da espera na fila. A dificuldade está justamente em encontrar a transformada $\overline{W}_{\overline{q}}(s)$, o que geralmente necessita de conceitos avançados da teoria de variáveis complexas.

3.2.4 Exemplo - $G/M/1$

Considerando a mesma amostra do exemplo em 3.1.2 para os tempos entre chegadas e tempos de atendimento, resolve-se testar a distribuição Gama para os tempos entre chegadas, ou seja, um modelo $G/M/1$.

Novamente, o teste de Kolmogorov-Smirnov foi realizado para testar a hipótese de que os tempos entre chegadas são distribuídos por Gama, obtendo-se o p-valor de 0.1502. Logo, não se rejeita que os tempos entre chegada são provenientes de uma distribuição Gama.

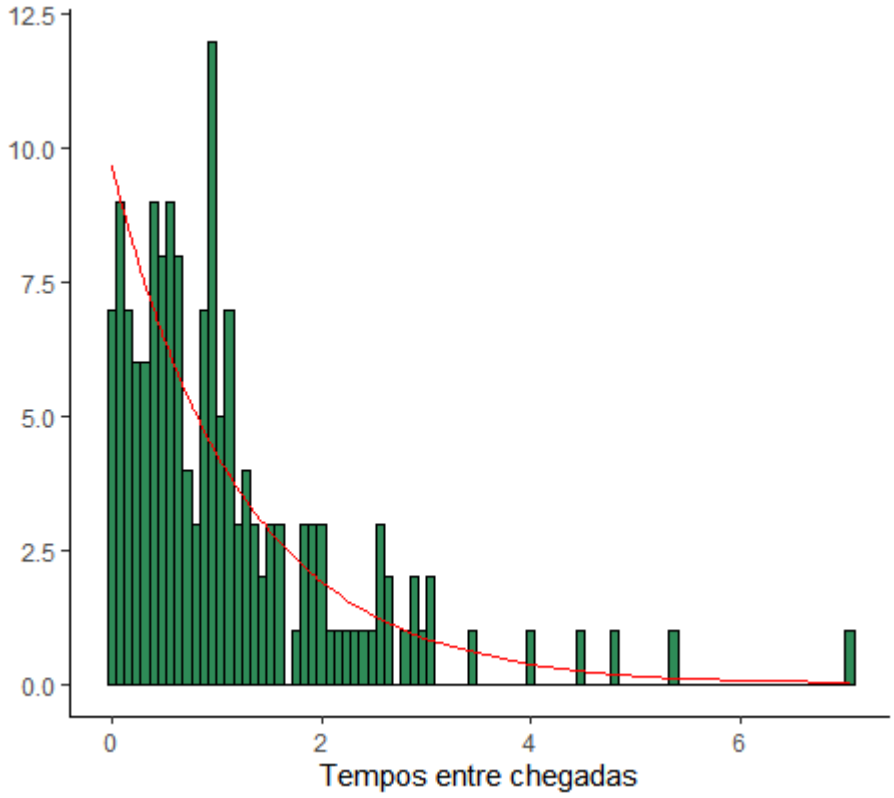


Figura 3: Histograma dos tempos entre chegada, com distribuição exata em vermelho.

As estimativas obtidas a partir da amostra são os mesmos valores do exemplo anterior, da Seção 3.1.2: $\lambda = 1.119$ e $\mu = 1.951$. Necessita-se apenas da estimativa de μ para encontrar numericamente a raiz de $z = A^*[\mu(1-z)]$, aproximada pela série $\sum_{i=1}^{k} a_i e^{-\mu t_i(1-z)}$, em que a_i são as probabilidades de cada tempo entre chegada t_i.

A raiz r_0 é o número z que faz $A^*(z) = z$. Dessa maneira, temos $r_0 = 0.524$. Podemos então obter as medidas de desempenho do sistema

Medida	Fórmula	Valor real	Valor estimado	I.C. (95%)
L	$\frac{r_0}{1-r_0}$	1	1.1	[0.876, 1.415]
L_q	$\frac{r_0^2}{1-r_0}$	0.5	0.577	[0.409, 0.829]
W	$\frac{1}{\mu(1-r_0)}$	1 min	1.077 min	[0.813, 1.43]
W_q	$\frac{r_0}{\mu(1-r_0)}$	0.5 min	0.564 min	[0.38, 0.838]
$P(N \geq 3)$	r_0^3	0.125	0.144	[0.102, 0.201]
$P(T_q > 2)$	$r_0 e^{-\mu(1-r_0)2}$	0.068	0.082	[0.04, 0.145]

Tabela 3: Medidas de desempenho para exemplo de um sistema $G/M/1$.

Os intervalos de confiança estão menores e as estimativas das medidas de desempenho estão mais próximas dos valores reais do que no exemplo $M/M/1$, como era de se esperar, tendo em vista que o modelo real envolve a distribuição Gama para os tempos de atendimento.

3.3 Procedimentos de aproximação

No caso de situações que não podem ser descritas analiticamente, ou seja, na maioria das situações práticas reais, há a possibilidade de utilizar resultados aproximados. Foram propostos alguns métodos de aproximação muito úteis ao longo da segunda metade do século passado.

3.3.1 Limites e Inequações

Marshall (1968) desenvolveu o seguinte resultado:

$$W_q = \frac{-E[I^2]}{2E[I]} - \frac{E[U^2]}{2E[U]},$$

em que I corresponde ao tamanho do período ocioso (em que os servidores não atendem clientes) e U é a diferença entre a v.a. dos tempos de serviço S e a v.a. dos tempos entre chegadas T, isto é, $U = S - T$. Tem-se, então

$$\begin{aligned} E[I] &= \frac{E[X]}{q_0} = \frac{P[\text{sistema estar vazio no momento de uma chegada}]}{q_o} \\ &= -\frac{E[U]}{q_0} = \frac{1/\lambda - 1/\mu}{q_0}. \end{aligned}$$

q_0 representa a probabilidade no estado estacionário de que um cliente ao chegar não encontre o sistema ocupado por ninguém, portanto, $q_0 \leq 1$. Dessa forma, $E[I] \geq 1/\lambda - 1/\mu$. Pode-se então chegar aos limites (Marshall (1968) e Marchal (1978)):

$$W_q \leq \frac{1}{2}\left(\frac{Var[S] + Var[T]}{1/\lambda - 1/\mu}\right),$$

$$W_q \geq \frac{E[T^2] - E[U^2]}{2E[U]},$$

que podem ser reescritos como

$$\frac{\lambda^2(\sigma_B^2 + 1/\mu^2 - 2/\mu\lambda)}{2\lambda(1-\rho)} \leq W_q \leq \frac{\lambda(\sigma_A^2 + \sigma_B^2)}{2(1-\rho)},$$

sendo ρ a razão da intensidade do tráfico λ/μ, para o caso de um único servidor; σ_A^2, a variância da distribuição dos tempos de serviço, e σ_B^2, a variância da distribuição dos tempos entre chegadas.

Essas inequações servem para todas as filas G/G/1, mas têm a restrição de que o limite inferior é positivo se, e somente se, $\sigma_B^2 > (2-\rho)/\lambda\mu$. Portanto, nem sempre são úteis.

Além disso, se as distribuições das chegadas e dos atendimentos são conhecidas, há outro limite inferior a considerar, dado por $W_q \geq r_0$, em que r_0 é a única raiz não negativa quando $\rho > 1$ de

$$f(z) = z - \int_{-z}^{\infty} [1 - U(t)]dt = 0,$$

e, assim, o intervalo fica

$$max\left(0, r_0, \frac{\lambda^2(\sigma_B^2 + 1/\mu^2 - 2/\mu\lambda)}{2\lambda(1-\rho)}\right) \leq W_q \leq \frac{\lambda(\sigma_A^2 + \sigma_B^2)}{2(1-\rho)}.$$

Os limites para W_q no caso mais geral G/G/c, em que $\rho = \lambda/c\mu$ são dados por (Gross e Harris, 1998):

$$max\left(0, \frac{\lambda^2\sigma_B^2 + c^2\rho(\rho-2)}{2\lambda c^2(1-\rho)} - \frac{\mu(c-1)(\sigma_B^2 + 1/\mu^2)}{2c}\right) \leq W_q \leq \frac{\lambda(c\sigma_A^2 + \sigma_B^2/c)}{2c(1-\rho)}.$$

3.3.2 Aproximações

Os métodos de aproximação pode ser divididos em três categorias (Bhat et al., 1979). O primeiro método utiliza limites e inequações para aproximar; o segundo utiliza um sistema de filas conhecido para aproximar os resultados de um sistema em estudo, sendo chamado de aproximação de sistemas; por exemplo, aproximar M/G/c por um M/E_k/c. O terceiro tipo de aproximação é a aproximação de processos, em que o próprio processo de filas é aproximado por um processo mais fácil de se lidar; um exemplo desse

último tipo é a substituição de um processo de filas discreto por um contínuo ou fluido.

Aproximação por Limites e Inequações

Baseado no fato de que o limite superior (visto acima) melhora quando $\rho \to 1$, faz sentido multiplicar esse limite por uma fração contendo ρ que aproxima ρ a um. Marchal (1978) propôs o quociente

$$\frac{\rho^2\sigma_A^2 + \sigma_B^2}{\sigma_A^2 + \sigma_B^2}$$

que leva à aproximação

$$\hat{W}_q = \frac{\rho(\lambda^2\sigma_A^2 + \mu^2\sigma_B^2)}{2\mu(1-\rho)}.$$

Esse resultado é exato para os modelos $M/G/1$ e $D/D/1$ e funciona bem para $G/M/1$. A performance da aproximação para o modelo geral $G/G/1$ deteriora quanto mais a distribuição dos tempos de serviço e a dos tempos entre chegadas desviam-se da exponencialidade. No entanto, sua capacidade de aproximação aumenta quanto maiores os valores de intensidade de tráfico, devido à natureza assintótica do limite superior.

A aproximação pode ser reescrita como produto de três termos: um fator de intensidade de tráfico, um fator de variabilidade, e um fator de escala de tempo,

$$\hat{W}_q = \left(\frac{\rho}{1-\rho}\right)\left(\frac{C_A^2 + C_B^2}{2}\right)\left(\frac{1}{\mu}\right),$$

em que C representa os respectivos coeficientes de variação.

Para o modelo $G/G/c$, há uma aproximação simples dada pela fórmula de Allen-Cunneen (AC) - (Allen, 1990):

$$\hat{W}_q = \frac{p_{cb}}{c(1-\rho)}\left(\frac{C_A^2 + C_B^2}{2}\right)\left(\frac{1}{\mu}\right),$$

em que p_{cb} é a probabilidade de que todos os servidores estejam ocupados em um sistema $M/M/c$. Então, sendo $r = \lambda/\mu$ e $\rho = r/c$, reescreve-se

$$\hat{W}_q = \frac{r^c P_0}{c \times c!(1-\rho)^2}\left(\frac{C_A^2 + C_B^2}{2}\right)\left(\frac{1}{\mu}\right).$$

Aproximação por Sistemas

Esse método consiste em aproximar um modelo por outro. Geralmente, aproximam-se modelos mais gerais por modelos Erlang, que possuem grande flexibilidade, como visto na seção 3.2.2. Aproxima-se, por exemplo, um modelo $M/G/c$ pelo mais simples $M/E_k/c$. Para tanto, uma aproximação da função do tempo de espera na fila, que vale para qualquer

modelo $GE/GE/1$, é dada por

$$W_q(t) = 1 + \sum_{i=1}^{n} k_i e^{z_i t},$$

em que $\{k_i\}$ representa a probabilidade de i chegadas em um tempo de serviço, e z_i são as raízes da equação polinomial $A^*(-s)B^*(s) - 1 = 0$.

Smith (1953) mostrou que um resultado similar é válido para modelos com tempos entre chegada arbitrários.

Aproximação por Processos

Muito útil para situações em que a intensidade de tráfico é muito próxima de 1 ($1 - \epsilon < \rho < 1$), denominada saturada ou de tráfico pesado. Os trabalhos de Smith e Wilkinson (1965) desenvolvem a discussão para soluções em problemas de tráfico pesado.

Para sistema de filas $G/G/c$ em tráfico pesado, tem-se a aproximação muito útil do tempo de espera na fila

$$W_q^{(H)} = \frac{1}{2}\frac{Var[T] + (1/c^2)Var[S]}{1/\lambda - 1/c\mu}.$$

3.4 Simulações

É muito comum não ser possível desenvolver modelos analíticos para sistemas de filas, o que é devido a características dos mecanismos de chegada e de serviço, da complexidade estrutural do sistema, da disciplina de filas ou combinações de todos esses fatores. Por exemplo, um sistema com múltiplos canais de atendimento e múltiplas fases de atendimento, em que o cliente pode retornar a determinada fase anterior e depois simplesmente sair do sistema, e os tempos de serviço são truncados normalmente distribuídos e com um sistema de prioridade complexo, é impossível de modelar analiticamente. É o caso de hospitais e de clínicas, em que há triagem, consulta, exame e reagendamento de consulta, seguindo a prioridade de atendimento prevista em lei, além das regras médicas sobre casos mais graves que devem ser atendidos prioritariamente.

Outrossim, mesmo os modelos mais gerais e abrangentes fornecem apenas resultados sob condição de estacionariedade. Se o pesquisador estiver interessado em efeitos transientes ou se as distribuições de probabilidade se alterassem com o tempo, provavelmente não seria possível desenvolver soluções analíticas, e nem mesmo soluções numéricas. Para problemas como esses, é necessário utilizar simulações (Gross e Harris, 1998).

Na análise via simulação, tem-se todos os problemas usuais associados à experimentação a fim de se fazer inferências acerca do mundo real, não prescindindo da preo-

cupação com certos aspectos como tempo de processamento, quantidade de replicações e significância estatística.

Outro ponto a se considerar em simulações é quando se quer encontrar a solução ótima para um sistema de filas. Suponhamos que se deseja determinar o número ótimo de canais de atendimento ou a taxa de serviço ótima em um sistema de custos conhecidos mesmo ao variar sua estrutura. Se um modelo analítico pode ser desenvolvido, a otimização matemática (cálculos diferenciais, por exemplo) pode ser feita. Entretanto, no caso de simulações, as técnicas de exploração dos resultados experimentais não são tão bem definidas e claras quanto as soluções matemáticas. É possível usar técnicas de otimização em conjunto com técnicas de simulação. Obviamente, a otimização em situações mais complexas é mais difícil do que em situações simples em que soluções analíticas podem ser encontradas.

Mas, novamente, em muitas situações, simulação é a única maneira de proceder, tendo encontrado muitos usos em transporte, indústria e manufatura, e comunicações. Tais sistemas são, comumente, estocásticos por natureza, com uma variedade de processos aleatórios interagindo de maneiras complexas. Sem simplificar suposições sobre a natureza da aleatoriedade, das distribuições de probabilidade, etc., modelagem analítica não é uma opção (Gross e Harris, 1998).

Elementos de um Modelo de Simulação

Há três elementos principais em modelagem por simulação (Gross e Harris, 1998): (1) construção do modelo gerador de dados; (2) rastreamento e contabilização de cada processo simulado ocorrido; (3) análise dos resultados.

Como o interesse é a modelagem de sistemas estocásticos, torna-se necessário selecionar e então gerar o fenômeno estocástico apropriado no computador. Por exemplo, o atendimento de emergência em um hopital consiste em uma rede de filas com diferentes distribuições de tempos entre chegadas e de tempos de serviço. Há fila para a recepção, para a triagem, para o atendimento médico, para exame, entre outras atividades. Devem-se decidir quais distribuições utilizar para representar essas estruturas de chegada e de atendimento. Então, gerações aleatórias dessas diferentes distribuições são feitas para que o sistema seja observado em funcionamento. Uma vez que as distribuições são escolhidas e gerados os dados, a fase de rastreamento e contabilização compila todas as movimentações pelo sistema para se poder calcular as medidas de performance. Na análise de resultados, serão computadas essas medidas e empregadas técnicas apropriadas para se testar e validar as performances do sistema.

3.4.1 Modelagem dos tempos do processo

Necessária em qualquer modelagem probabilística, incluindo analítica e numérica, a modelagem dos tempos do processo pode ser dividida em dois problemas principais: seleção das distribuições e estimação dos parâmetros da distribuição escolhida.

Porém, há situações em que é recomendável procurar distribuições empíricas, ao invés de procurar por uma distribuição teórica que se encaixe - Fox (1981) e Kelton (1984). Em casos que haja muitos registros históricos e a distribuição é complexa, por exemplo, é preferível adotar um processo não paramétrico de reamostragem, no qual observações sintéticas são obtidas sem a especificação da forma funcional das distribuições.

Estimação dos parâmetros

Após escolher as distribuições para modelar o problema e assumindo que há acesso aos dados reais de tempos entre chegadas e tempos de serviço/atendimento, tem-se uma amostra aleatória de tamanho n, denotada por $t_1, t_2, \ldots, t_n$.

Dois métodos clássicos de estimação de parâmetros são o método de máxima verossimilhança (MLE) e o método dos momentos (MOM). O primeiro consiste em escrever a função de verossimilhança, que é a densidade conjunta da amostra, se os elementos que a compõem forem independentes: $L(\theta) = \prod_{i=1}^{n} f(x_i)$, e então maximizar o valor do parâmetro θ para se encontrar o estimador de máxima verossimilhança, sendo $f(x)$ a função de densidade.

O método dos momentos simplesmente iguala os momento empíricos e teóricos, resolvendo a equação obtida para o parâmetro θ.

O MLE reconhecidamente possui propriedades muito boas para seus estimadores, como consistência ($\lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq k) = 0, \forall\ k$), suficiência ($f_{X|\hat{\theta}}$ independe de θ), e ser estimador de mínima variância não-viesado.

Porém, há casos em que é mais fácil obter estimadores por MOM. Por exemplo, a fórmula para o modelo $M/G/1$ depende apenas da média e da variância da distribuição do tempo de serviço. O limite superior de Kingman-Marshall e a aproximação para tráfico pesado dependem apenas dos dois primeiros momentos (média e variância) das distribuições dos tempos entre chegada e dos tempos de serviço.

Além disso, se a média e a variância encontradas utilizando MLE forem muito diferentes dos valores da média e da variância dos dados, optaria-se por MOM ou outro método de estimação paramétrica (Gross e Harris, 1998).

Para maiores detalhes acerca desses métodos, de ampla utilidade e relevância, recomendamos a leitura de um livro introdutório de Estatística.

Seleção das Distribuições

É complicado decidir a distribuição adequada para representar os tempos do processo (tempos de serviço e de entre chegada), porque depende de amplo conhecimento sobre as caracetrísticas de diversas distribuições e o entendimento de cada processo que ocorre na situação analisada e do próprio processo como um todo. Por exemplo, se há grande variabilidade entre os tempos de atendimento, a exponencial deve servir. Algumas medidas auxiliam nessa decisão.

O coeficiente de variação ($CV = \sigma/\mu$, ou razão do desvio padrão em relação à média) pode ser muito útil, pois possibilita comparar os $CV's$ de várias distribuições. A exponencial, por exemplo, possui $CV = 1$; Erlang tipo k, $k > 1$, tem $CV < 1$; e a hiperexponencial ($k > 1$), possui $CV > 1$.

A função de taxa de falha (*hazard function*) é outro meio que oferece um *insight* geral sobre a distribuição adequada. Fornece uma probabilidade aproximada de que o atendimento a um cliente termine dado que já está em atendimento a um tempo t, e é dada por

$$h(t) = \frac{f(t)}{1 - F(t)},$$

e pode ser crescente em t, decrescente em t, constante ou uma combinação. O caso constante implica a propriedade de perda de memória.

Caso se acredite que o processo de atendimento na situação analisada compreende que quanto mais tempo um cliente esteve em atendimento, maior a probabilidade de que o serviço seja completado no próximo tempo t, então deseja-se uma distribuição $f(t)$ para a qual $h(t)$ seja crescente, ou seja, que a taxa de falha seja crescente. Vemos, portanto, que a taxa de falha é outro meio importante para selecionar as distribuições candidatas a representarem os tempos do processo.

Suponha que se deseja uma condição também crescente para a taxa de falha, mas que essa taxa de crescimento acelere de acordo com t. A Weibull pode ser uma boa candidata, em que $F(t) = 1 - e^{\beta t^{\alpha}}$. Na verdade, dependendo do valor do parâmetro de forma α, podem-se até mesmo obter taxas de falha decrescentes.

Barlow e Proschan (1965) provaram que todas as ditribuições com taxa de falha crescente possuem $C < 1$ e todas as distribuições com taxa de falha decrescente possuem $C > 1$, sendo que o inverso não é verdadeiro. Assim, se é observado que a condição de taxa de falha crescente é consistente com os tempos observados, devemos procurar uma distribuição com $C < 1$.

É importante verificar se a amostra é IID. Os trabalhos de Leemis (1996) e de Law e Kelton (1991) são interessantes para testar tal hipótese.

Assumindo que se tem uma amostra IID, um histograma dos dados também pode ser bastante útil, levando em conta que o formato do histograma depende do comprimento dos

intervalos utilizados para calcular as frequências, que é a quantidade de observações que caem em determinado intervalo. Uma regra prática é escolher comprimentos de intervalos de forma que haja pelo menos cinco observações em cada intervalo e que haja pelo menos cinco intervalos, mas devem-se experimentar vários tamanhos de intervalo (Gross e Harris, 1998).

Testando a distribuição candidata

Após observar o histograma dos dados, considerando as características da estrutura do sistema que se está investigando e das distribuições, escolhe-se uma distribuição candidata. Existe uma variedade de testes estatísticos que podem ser realizados a fim de verificar se a distribuição candidata é uma escolha razoável.

O mais comum é o teste χ^2 de qualidade de ajuste. Outro teste comum é o teste de Kolmogorov-Smirnov (KS) , que compara os desvios da f.d.a. empírica com os da f.d.a. teórica, sendo sua estatística um desvio absoluto máximo modificado. Uma variação do teste KS é o teste de Anderson-Darling, que utiliza não apenas o desvio máximo como KS, mas todos os desvios, fazendo uma média ponderada do quadrado dos desvios, com os pesos sendo as maiores caudas da distribuição. Um outro teste é o teste F. Para a exponencial, geralmente é o mais poderoso e de melhor performance que os outros testes.

Diversos softwares, como o próprio R (R Core Team 2019), em funções como *geom_smooth*, recomendam uma distribuição para certo conjunto de dados de acordo com vários critérios além desses testes estatísticos. Mas é preciso cautela, principalmente com relação aos momentos. Caso os momentos da distribuição teórica forem muito diversos dos obtidos pela amostra, o melhor é rejeitar tal distribuição. Juttijudata (1996) e Gross e Juttijudata (1997) demonstraram quão importantes o primeiro e o segundo momento são. Para a aproximação de casos de tráfico pesado, por exemplo, esses momentos são a única coisa que importa, sendo que a distribuição nem entra nas fórmulas.

É recomendável, em alguns casos, simplesmente utilizar a distribuição empírica dos dados, ao invés de procurar por uma distribuição teórica que se encaixe - Fox (1981) e Kelton (1984).

Geralmente é preferível utilizar testes por simulação, que não dependem de resultados assintóticos, como os estudados no livro *Statistical Computing with R* (Rizzo, M. L., 2007).

Geração de números pseudo-aleatórios

Muitos programas contêm funções que geram números pseudo-aleatórios. "Pseudo" porque são completamente reproduzíveis por um algoritmo matemático e "aleatório"no sentido de que foram testados estatisticamente em relação à probabilidade equivalente de seleção de todos os valores e à independência estatística. No R, temos as funções

amplamente conhecidas *rnorm*, *rexp*, *runif*, *rpois*, ...

Existem vários métodos para se gerar observações representativas de qualquer distribuição de probabilidade, com f.d.a. $F(x)$. O mais popular é o da transformada inversa, que consiste em, primeiramente, gerar números aleatórios da Uniforme (0, 1), $r_1, r_2, \ldots, r_n$, e, então, igualá-los à fórmula da f.d.a. da distribuição que se deseja e encontrar os valores $x_1, x_2, \ldots, x_n$ correspondentes.

Para casos mais complexos, há uma variedade de técnicas alternativas, incluindo aceitação/rejeição, Monte Carlo Markov Chain (MCMC) e Sequential Monte Carlo (SMC).

Portanto, é possível gerar amostras aleatórias de qualquer distribuição para se proceder com os estudos computacionais.

Rastreamento e compilação do processo

Existe uma grande variedade de linguagens e pacotes para criar modelos de simulação. No R, por exemplo, os pacotes *queuecomputer* e *simmer* oferecem muita praticidade para a modelagem. Porém, ao utilizar funcionalidades prontas perde-se muita flexibilidade na construção do modelo.

Utilizar uma linguagem geral, como o C++ ou o próprio R, para modelar um processo de filas é complicado, pois requer definir distribuições para os tempos de espera e de atendimento em cada fase do processo, rastrear o caminho percorrido por cada "cliente" simulado, em que lugar ele está em momentos específicos, verificando se houve desistência ou se ele retornou a outro ponto do processo. Além disso, há que realizar todos os cálculos para as medidas de performance e verificar a adequabilidade delas e do modelo construído.

Ainda mais fáceis de utilizar que pacotes prontos de simulação, existem os chamados simuladores, que requerem pouca ou nenhuma programação, em que o usuário pode construir um modelo selecionando as opções que deseja de distribuição dos tempos, de número de clientes, etc., em uma plataforma com botões ou em que se arrastam ícones para desenhar o modelo. Logicamente, não há flexibilidade alguma na modelagem do processo. Alguns exemplos são SIMFACTORY, ProModel e NETWORK.

3.4.2 Análise dos resultados

Ao simular sistemas estocásticos e rodar o programa, são obtidos valores que precisam ser analisados para se chegar a concusões válidas.

Há dois tipos principais de modelos de simulação: finitos no tempo ou contínuos (Gross e Harris, 1998). Por exemplo, um banco que abre às 11h da manhã e encerra as atividades às 17h da tarde é um modelo limitado no tempo, que possui um momento de começo e um momento final; uma linha de montagem em que os trabalhadores continuam

o processo de onde pararam no turno anterior é um processo contínuo. Nesse último caso, resultados de estado estacionário geralmente são de interesse.

Em relação ao primeiro caso, dos processos limitados por momentos bem definidos de começo do processo e de finalização, é preciso replicar o experimento para se obter amostras das medidas obtidas e então calcular resultados. Ou seja, utilizando sequências de números aleatórios diferentes (a semente deve ser diferente) para se obter tempos de espera e de atendimento diversos cada vez que se rodar o simulador, é obtida uma amostra de observações independentes em que se podem aplicar métodos estatísticos clássicos.

Para n replicações, obtêm-se n valores para o tempo de espera máximo, $m_1, m_2, \ldots, m_n$. Sendo n suficientemente grande, o Teorema do Limite Central pode ser aplicado para obter um intervalo de confiança (IC) $100(1-\alpha)\%$, calculando-se a média e o desvio padrão por

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}$$

e

$$s_m = \sqrt{\frac{\sum_{i=1}^{n}(m_i - \bar{m})^2}{n-1}},$$

e então obter o IC para o tempo máximo de espera m,

$$\left[\bar{m} - \frac{t_{(n-1,1-\alpha/2)}s_m}{\sqrt{n}},\ \bar{m} + \frac{t_{(n-1,1-\alpha/2)}s_m}{\sqrt{n}}\right],$$

em que $t_{(n-1,1-\alpha/2)}$ é o valor crítico da distribuição t com $n-1$ graus de liberdade.

Para simulações contínuas em que se desejam obter resultados estacionários, sabe-se da teoria ergódica de processos estocásticos que

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T X^n(t)dt = E[X^n],$$

de forma que, se o programa for rodado por tempo suficiente, obtém-se um resultado próximo ao valor médio no limite. Mas não é claro quanto tempo é o suficiente.

Suponha que se deseja rodar uma simulação com n clientes e medir o tempo que um cliente espera em uma fila específica do processo. São obtidos, dessa forma, n tempos de espera, $w_1, w_2, \ldots w_n$. Simplesmente calcular o intervalo de confiança não funciona, porque a variância seria muito subestimada, uma vez que esses tempos de espera w_i são correlacionados.

Há procedimentos para estimar as correlações e obter uma estimativa do desvio padrão desses dados correlacionados, o que requer muito trabalho de estimação baseado apenas em um conjunto de dados, diminuindo a precisão estatística.

Para contornar o problema da correlação, a simulação é replicada m vezes, utilizando uma semente (*seed*) diferente para cada, como no caso de um processo limitado por horário. Cada vez que se roda a simulação, calcula-se a média $\bar{w}_j$ para cada j-ésima replicação,

$$\bar{w}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n},$$

em que w_{ij} é o tempo de espera do i-ésimo cliente na j-ésima replicação, $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, m$.

Agora, os $\bar{w}_j$ são independentes e pode-se calcular o IC de forma análoga:

$$\bar{w} = \frac{\sum_{j=1}^{m} \bar{w}_j}{m}$$

e

$$s_{\bar{w}_j} = \sqrt{\frac{\sum_{j=1}^{m} (\bar{w}_j - \bar{w})^2}{m-1}},$$

e então o IC torna-se

$$\left[\bar{w} - \frac{t_{(m-1,1-\alpha/2)} s_{\bar{w}_j}}{\sqrt{m}},\ \bar{w} + \frac{t_{(m-1,1-\alpha/2)} s_{\bar{w}_j}}{\sqrt{m}}\right],$$

em que $t_{(m-1,1-\alpha/2)}$ é o valor crítico da distribuição t com $m-1$ graus de liberdade.

3.4.3 Validação do Modelo

Validação está associada a quão precisa é a representação da realidade por um modelo, o que envolve a verificação do programa utilizado, para se ter certeza de que ele faz o que é pedido, e comparação com os resultados de outros programas de simulação ou de modelos analíticos, se possível.

O modelo pode ser rodado com uma variedade de condições, alterando a estrutura e os parâmetros, e, se em diversas simulações os resultados forem semelhantes, a validação do modelo é confirmada.

3.4.4 Exemplo - Simulação

Para esta parte, utilizamos as funções *queue_step* e *summary* do pacote do R "queuecomputer" para a resolução dos exemplos anteriores do modelo básico e do modelo geral nas seções 3.1.2 e 3.2.4, ainda utilizando a mesma amostra.

Medida	Valor real	Valor estimado	I.C. (95%)
L	1	1	[0.908, 1.09]
L_q	0.5	0.548	[0.475, 0.621]
W	1 min	1.133 min	[0.926, 1.34]
W_q	0.5 min	0.62 min	[0.434, 0.806]
$P(N \geq 3)$	0.125	0.119	-
$P(T_q > 2)$	0.068	0.12	-

Tabela 4: Medidas de desempenho calculadas por simulação.

A simulação levou 0.01 segundos, rodando em uma máquina Intel Core i5-7200U (velocidade de até 3.1 GHz) com memória RAM de 8 GB.

Por simulação, as estimativas calculadas para L e L_q são mais verossímeis do que utilizando os modelos analíticos, como vimos nos exemplos 3.1.2 e 3.2.4. Porém, para a estimação dos tempos de espera, W e W_q, os valores estimados pela simulação estavam mais distantes do que as estimativas do modelo analítico $G/M/1$.

Isso se deve ao fato de que o programa apenas calcula os tempos de saída de acordo com os tempos de chegada e tempos de serviço fornecidos, não levando em conta a suposição do modelo $G/M/1$ de que os tempos entre chegadas possuem distribuição Gama, informação incluída na função $A^*(z) = A^*[\mu(1-z)]$ (Seção 3.2.4). Sem essa informação adicional, era de se esperar que as estimativas dos tempos fossem menos acertadas.

3.4.5 Exemplo - Sistema com múltiplos níveis de atendimento

Considerando ainda o exemplo do hospital oftalmológico, mas sendo que, agora, o interesse é estudar o comportamento não apenas do atendimento na recepção, mas também do atendimento na enfermagem e do atendimento médico. Ou seja, há três processos de espera e de serviço dentro do sistema, sendo que um depende do outro. Além disso, há prioridade no atendimento na recepção a pessoas idosas. Impossível analisar um sistema como este analiticamente.

O sistema foi simulado com o pacote *simmer* (código no Apêndice, Seção 7.1). Os tempos entre chegadas dos pacientes e os tempos de serviço foram gerados a partir de distribuições normais: tempos entre chegadas, N(3,1); tempos de atendimento na recepção, N(3,1); tempos de atendimento na enfermagem, N(10,2); e tempos de atendimento médico, N(15,4).

A amostra final ficou composta por 57 pacientes, dentre os quais 3 idosos, e o conjunto de servidores, por 3 recepcionistas, 4 enfermeiros/enfermeiras e 5 médicos/médicas. Foram obtidos os valores $W = 34.765$ min e $W_q = 6.78$ min.

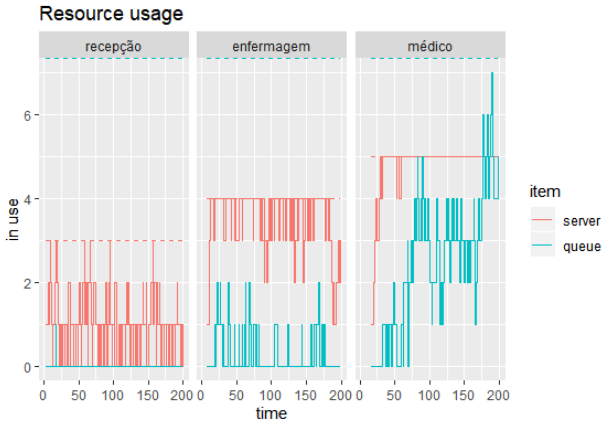
O pacote permite a visualização de como foram utilizados os postos de atendimento ao longo do tempo, os momentos em que se formaram filas e as taxas de utilização.

Na Figura 4, a primeira coluna com os gráficos (a), (c), e (e) representa a estrutura "original" de atendimento do hospital, com 3 recepcionistas, 4 enfermeiros/enfermeiras e 5 médicos/médicas. Ainda na figura 4, a segunda coluna com os gráficos (b), (d), e (f) mostra uma nova estrutura de atendimento do hospital, com apenas 1 recepcionista, 3 enfermeiros/enfermeiras e 5 médicos/médicas.

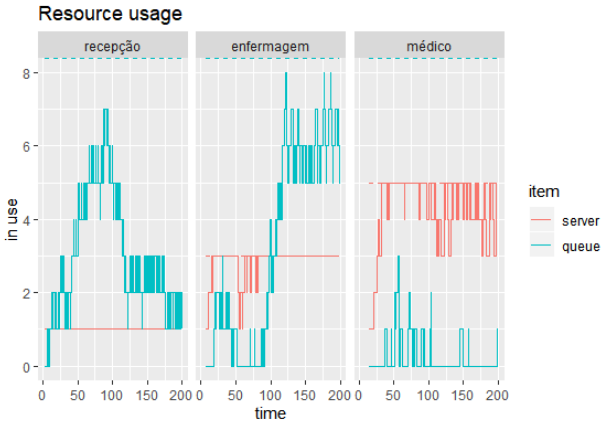
Com a diminuição da quantidade de recepcionistas e profissionais de enfermagem, observa-se que há formação de filas na recepção e na enfermagem, sem filas no atendimento médico, sendo que na estrutura "original" havia formação de filas apenas no atendimento médico.

Antes havia um subaproveitamento na recepção, com apenas cerca de 36% de taxa de utilização. Com a alteração da quantidade de funcionários, o aproveitamento do trabalho na recepção foi para 100%. Porém, a espera na fila quase triplicou, subindo para $W'_q = 18.64$ min, e o tempo total do usuário no sistema também subiu, resultando em $W' = 46.25$ min.

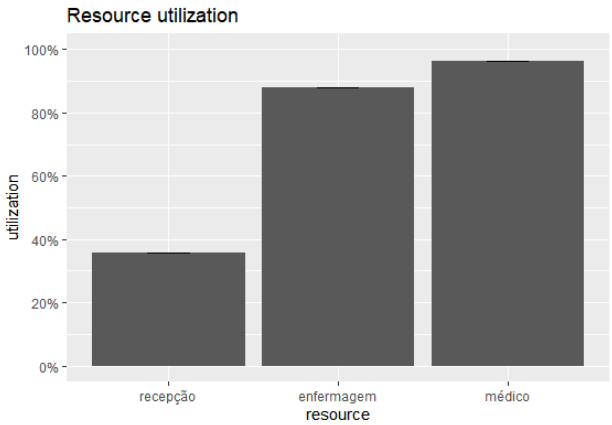
Este exemplo ressalta a importância e grande utilidade de se estudar um sistema a partir de um modelo simulado, pois permite identificar problemas e avaliar alterações que possam melhorar a experiência dos usuários e/ou aproveitar os recursos disponíveis de forma eficiente.



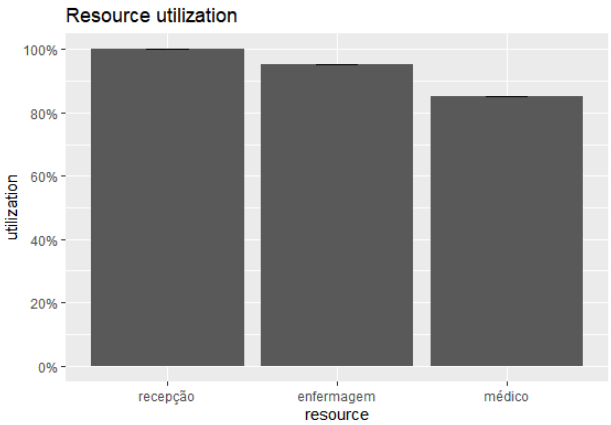
(a) Utilização e filas - situação 1.



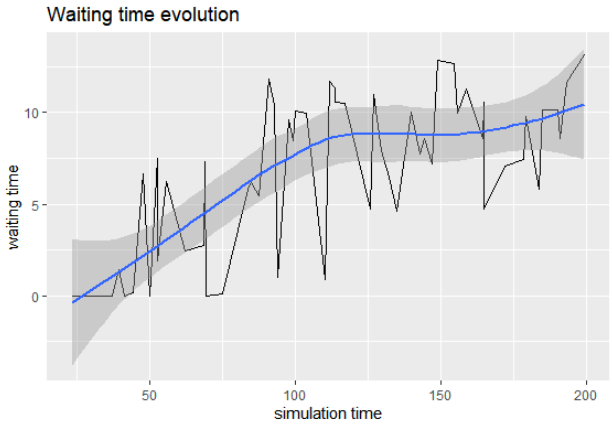
(b) Utilização e filas - situação 2.



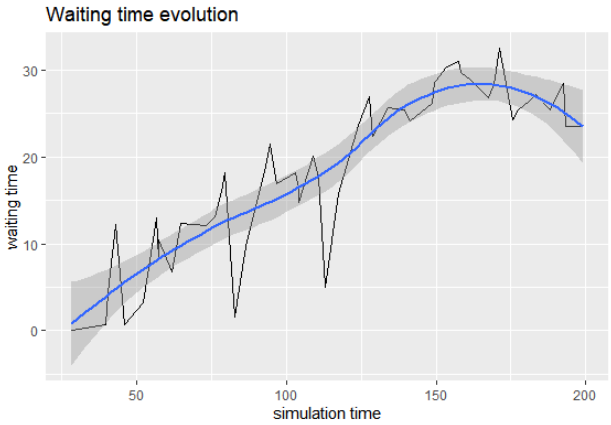
(c) Taxas de utilização - situação 1.



(d) Taxas de utilização - situação 2.



(e) Evolução do tempo de espera - situação 1.



(f) Evolução do tempo de espera - situação 2.

Figura 4: Utilização e espera de um exemplo de atendimento em um hospital. (a) e (b): Quantidade de servidores (em vermelho) e tamanho da fila (em azul) ao longo do tempo; (c) e (d) taxas de utilização dos postos de atendimento; (e) e (f) tempo de espera do paciente ao longo do tempo com ajuste de uma banda de confiança pelo método *loess*.

4 Metodologia

Como o projeto busca descrever e investigar o processo de atendimento de emergência hospitalar, faz-se necessária a descrição da fila e das características do processo (padrão de chegada dos pacientes; padrões de serviço; disciplina da fila; capacidade do sistema; número de canais de serviço; estágios de serviço).

Foi feita uma fundamentação teórica da Teoria de Filas. Devido à complexidade de processos como o descrito na Figura 5, a estimação dos parâmetros do modelo não foi conduzida analiticamente, mas sim com a implementação de simulações computacionais, implementadas no software livre R (R Core Team 2019).

Embora existam alguns pacotes (conjunto de funções já implementadas) de sistemas de filas que fornecem funcionalidades interessantes (como *queuecomputer* e *simmer*), foi desenvolvido um programa próprio, para que se pudesse modelar com maior liberdade, obtendo melhores resultados tanto em termos da estimação como da visualização. Além disso, foi desenvolvido um aplicativo básico com a interface Shiny, para facilitar a utilização pelo usuário, que pode facilmente alterar parâmetros, como número de canais de atendimento, e visualizar os resultados.

Desta forma, foram investigados diferentes sistemas de atendimento, variando-se a quantidade de médicos, enfermeiros, recepcionistas em cada fase do processo e alterando o horário da troca de turnos.

4.1 Descrição probabilística do processo

As taxas de chegada dos pacientes, assim como a taxa dos tempos de atendimento em cada estágio do processo, são os parâmetros de interesse a serem estimados, a partir dos quais todo o sistema de atendimento pode ser simulado.

Seja X uma variável aleatória que possui distribuição Hiperexponencial. Então,

$$f_X(x) = \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i x}.$$

Ou seja, é uma mistura de k exponenciais, sendo p_i a probabilidade de que X assuma uma exponencial com taxa λ_i.

Dessa maneira, temos

$$
\begin{aligned}
&T_r;\\
&S_r \sim Hiperexp(\lambda_r, p_r);\\
&C_r \sim Multinomial(\tilde{p});\\
&\tilde{p_r} \sim Dirichlet(\alpha_{id}, \alpha_{no});\\
&S_t \sim Hiperexp(\lambda_t, p_t);\\
&C_t \sim Multinomial(\tilde{p});\\
&\tilde{p_t} \sim Dirichlet(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5);\\
&S_c \sim Hiperexp(\lambda_c);
\end{aligned}
$$

T_r é a distribuição empírica dos tempos de chegada à recepção, feita por reamostragem dos dados referentes à variável $DH\_RETIRADA\_SENHA$, conforme descrito na próxima seção sobre os materiais utilizados no trabalho. S_r é a distribuição dos tempos de atendimento na recepção. C_r é a distribuição que descreve a prioridade de atendimento na recepção, cujos hiperparâmetros são $\{p_{id}, p_{no}\} = \tilde{p}$, representando as probabilidades do paciente ser idoso e de não ser, com hiperparâmetros de uma Dirichlet $(\alpha_{id}, \alpha_{no})$.

Para a triagem, temos um modelo análogo, com a diferença de que não há distribuição para os tempos de chegada a esse estágio do atendimento na emergência, pois utilizamos as estimativas dos tempos de saída do atendimento na recepção como os tempos de chegada à triagem, havendo apenas a distribuição dos tempos de atendimento na triagem, S_t . Além disso, a classificação de risco feita na triagem possui 5 categorias de risco, cujas estimativas $\{p_1, p_2, p_3, p_4, p_5\} = \tilde{p_t}$ são as probabilidades dos níveis de risco do mais urgente (1) ao menos urgente (5), com hiperparâmetros de uma Dirichlet $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$.

O tempo de atendimento nos consultórios médicos será descrito por S_c, com taxa λ_c e a partir desse vetor são selecionados os tempos de atendimento na sala amarela e na sala vermelha, para onde vão os pacientes classificados nos níveis de gravidade 2 e 1, respectivamente.

4.2 Materiais

O conjunto de dados fornecido para este trabalho é composto por 81070 observações e 16 variáveis sobre o atendimento emergencial no Hospital das Forças Armadas de Brasília. As cinco primeiras observações estão apresentadas a seguir:

	CD_ATENDIMENTO	CD_PACIENTE	DH_RETIRADA_SENHA	DH_CADASTRO
1	XXXXX	98870	01/01/19 00:27	01/01/19 00:29
2	XXXXX	51336	01/01/19 01:21	01/01/19 01:21
3	XXXXX	82989	01/01/19 01:46	01/01/19 02:45
4	XXXXX	45586	01/01/19 03:20	01/01/19 03:20
5	XXXXX	83876	01/01/19 04:33	01/01/19 04:34

	DH_INI_CLASSIF	DH_FIM_CLASSIF	DH_INI_ATEND_MED	DH_FIM_ATEND_MED
1			01/01/19 01:00	01/01/19 01:56
2			01/01/19 01:29	01/01/19 01:36
3			01/01/19 03:10	01/01/19 03:15
4			01/01/19 03:28	01/01/19 03:50
5			01/01/19 05:02	01/01/19 11:46

	DH_ENT_SL_AM	DH_SAIDA_SL_AM	DH_ENT_SL_VERM	DH_SAIDA_SL_VERM
1				
2				
3				
4				
5	01/01/2019 08:01	01/01/2019 11:01		

	DH_ALTA_MEDICA	DH_ADMISSAO	ESPECIALIDADE	DIA_DA_SEMANA
1	01/01/19 01:56		RADIOLOGIA	TERÇA-FEIRA
2			RADIOLOGIA	TERÇA-FEIRA
3	01/01/19 03:15		ORTOPEDIA/TRAUMATOLOGIA	TERÇA-FEIRA
4			CLINICA GERAL	TERÇA-FEIRA
5	01/01/19 11:46	01/01/19 12:06	CIRURGIA GERAL	TERÇA-FEIRA

As duas primeiras colunas são meramente o cadastro e a identificação numérica dos pacientes. O cadastro é realizado na recepção, onde trabalham dois recepcionistas no turno do dia, de 7h às 19h, e apenas um recepcionista no turno da noite, de 19h às 7h.

As colunas 3 a 14 fornecem informações da data e do horário em que os pacientes passaram em cada estágio do atendimento hospitalar, desde a recepção até a possível internação.

A penúltima coluna informa a especialidade médica a que foram transferidos os pacientes, variável com 30 categorias; se a especialidade a que foi transferido um paciente é apenas "clínica geral", quer dizer que ele não foi encaminhado para outro departamento do hospital.

A última coluna informa apenas o dia da semana em que o paciente deu entrada na emergência do hospital.

Também foram fornecidas mais informações sobre a quantidade de funcionários trabalhando na emergência do hospital. Na triagem, há apenas um enfermeiro. Há oito clínicos, sendo que dois atendem nos consultórios, três na sala amarela (para onde vão os pacientes com casos de segunda maior urgência), dois na sala vermelha (para onde vão os pacientes de maior urgência) e um fica em revezamento.

A descrição completa de cada variável encontra-se no Apêndice, Seção 7.2.

A Figura 5 abaixo é o fluxo completo do atendimento na emergência do Hospital das Forças Armadas de Brasília, fornecido pelo projeto HFA Data & Care.

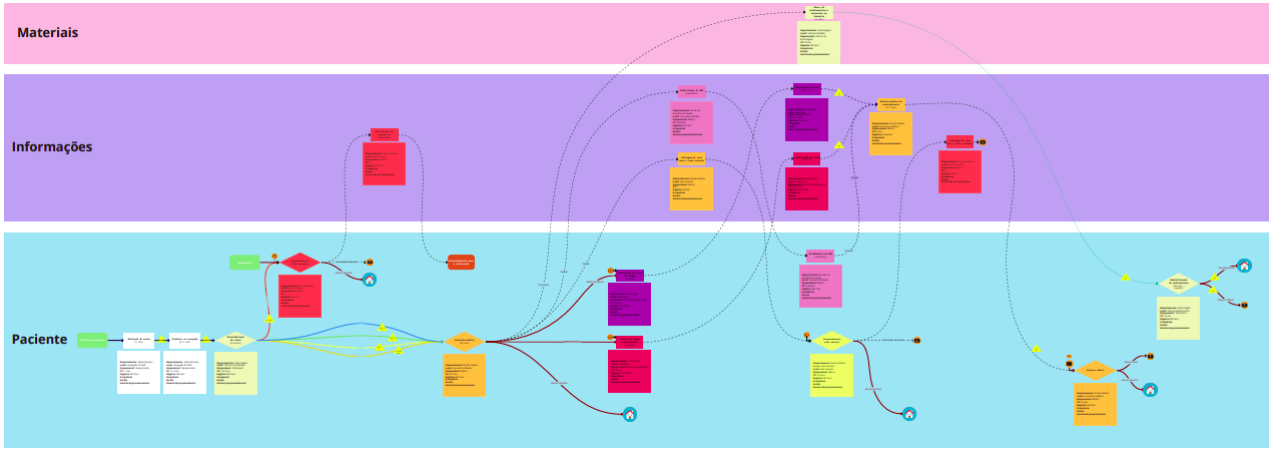


Figura 5: Fluxo de atendimento na emergência do HFA de Brasília (HFA Data & Care).

4.3 Algoritmo QDC

O processo de simulação implementado neste trabalho é baseado no algoritmo conhecido como QDC - *queue departure computation* - que registra os tempos de saída do atendimento de cada usuário, que são também os tempos em que ficam disponíveis os canais de atendimento, e então aloca o próximo usuário de acordo com o menor tempo de disponibilidade dos canais, ou seja, aloca o próximo usuário para o primeiro funcionário que estiver livre.

Considerando um sistema de atendimento *FCFS*, há uma única fila atendida por um número fixo de servidores K. O i-ésimo usuário/cliente seleciona o primeiro servidor disponível $p_i = argmin(b_i)$ de $b_i = \{b_{ik} | k \in 1 : K\}$, que é o vetor de tamanho K que representa os tempos em que cada servidor estará livre. Dessa forma, o tempo de saída do i-ésimo cliente é dado por $d_i = max(a_i, b_{pi}) + s_i$, sendo s_i o tempo de atendimento do i-ésimo cliente.

A chave do funcionamento do algoritmo é a alocação de cada cliente a um servidor de acordo com o vetor $\mathbf{b}$, que é atualizado continuamente, representando o estado do sistema.

Portanto, precisa-se apenas do vetor $\mathbf{a}$, representando os tempos de chegada (ordenados) dos usuários ao sistema, e do vetor $\mathbf{s}$, para simular o processo de espera e atendimento desse sistema.

Algoritmo QDC para K fixo

function($\mathbf{a} \in \mathbb{R}_+^n$, $\mathbf{s} \in \mathbb{R}_+^n$, $\mathbf{K} \in \mathbb{N}$)
  Ordenar ($\mathbf{a}$, $\mathbf{s}$) ascendentemente em termos de $\mathbf{a}$
  Criar vetor $\mathbf{p} \in \mathbb{N}^n$
  Criar vetor $\mathbf{b} \in \mathbb{R}_+^K$
  Criar vetor $\mathbf{d} \in \mathbb{R}_+^n$
  $b_k \longleftarrow 0, \forall\ k \in 1 : \mathbf{K}$
  for $i \in 1 : n$ **do**
    $p_i \longleftarrow \text{argmin}(\mathbf{b})$
    $b_{p_i} \longleftarrow max(a_i, b_{pi}) + s_i$
    $d_i \longleftarrow b_{p_i}$
  end for
  return ($\mathbf{d}$, $\mathbf{p}$)
  end function

Esse algoritmo pode simular qualquer fila $G/G/K/\infty/FCFS$, em que K pode ser escolhido como arbitrariamente grande, as distribuições entre chegadas e de serviço podem ser completamente gerais e até mesmo possuir uma estrutura de dependência entre elas (Ebert et al, 2019).

O simulador criado para este trabalho tem por base o algoritmo QDC, com distribuições hiperexponenciais para os tempos de chegada na recepção e para cada tempo de atendimento, sendo que os tempos de chegada em um estágio são os tempos de saída do estágio anterior. A distribuição Hiperexponencial foi esolhida para representar os tempos do sistema de acordo com a indicação dada pelo coeficiente de variação (Seção 3.4.1) $CV = \sigma/\mu > 1$ para os dados disponíveis dos horários de atendimento ($CV = 2.49$ para o atendimento médico).

Foi feita a adição de algumas funcionalidades:

- A quantidade K de servidores pode ser alterada ao longo do tempo;
- A estrutura de atendimento comporta prioridade.

Assim, pode-se modelar em função de turnos de trabalho e da classificação dos pacientes segundo algum critério. Os critérios de prioridade utilizados foram idoso ou não para atendimento na recepção e na triagem, e os cinco graus de classificação de risco de acordo com a gravidade da emergência médica, utilizados após a triagem, e de acordo com os quais o paciente vai para os consultórios regulares de atendimento médico (níveis 3 a 5 de emergência) ou para a sala amarela (nível 2 de emergência) ou para a sala vermelha (nível 1 de emergência).

Deve-se observar que a estrutura de atendimento por prioridade simulada pelo algoritmo funciona de maneira aproximada se o tempo de espera na fila for muito grande, pois, nesse caso, o algoritmo não é capaz de priorizar o atendimento de algúem que chegue na fila em um momento muito posterior ao tempo mínimo em que o atendimento estará disponível. Nesse caso, o algoritmo irá identificar esse paciente com prioridade em um

momento posterior, tornando o seu tempo de espera maior do que deveria ser. Ou seja, há um leve viés positivo no tempo simulado de espera de pacientes com prioridade caso o tempo de espera na fila seja demasiado grande.

O modelo é iterado $m = 100$ vezes, sendo os valores obtidos para as medidas de desempenho em cada iteração apresentados na Seção 5.5, conforme descrito na Seção anterior 3.4.2.

A quantidade n de pacientes que chegam diariamente à emergência do hospital para serem atendidos varia a cada iteração, sendo sorteado (sem reposição) um valor da quantidade de pacientes atendidos por dia em cada um dos 365 dias do ano.

4.4 Validação do programa implementado

Como forma de validar o funcionamento do programa implementado neste trabalho, foi realizada uma simulação com um pacote pronto do R e uma simulação com os mesmos dados no programa desenvolvido.

4.4.1 Comparação com *queuecomputer*

Para demonstrar a validade do algoritmo utilizado no modelo implementado considerou-se uma fila $M/M/2/\infty/FCFS$, com $n = 10000$ usuários, uma vez que, se o algoritmo é válido para um sistema $M/M/K$, então ele é válido para qualquer $G/G/K$, o que se deve ao fato de que qualquer (**a**, **s**) não nulo pode ser originado de duas distribuições exponenciais, mesmo que a probabilidade de uma determinada ocorrência seja extremamente pequena (Ebert et al, 2019).

Os dados utilizados para os tempos de chegada **a** e para os tempos de serviço **s** são gerados aleatoriamente.

```
R> set.seed(2021)
R> n_customers <- 10^4
R> lambda_a <- 1/2
R> lambda_s <- 1/1.5
R> interarrivals <- rexp(n_customers, lambda_a)
R> arrivals <- cumsum(interarrivals)
R> service <- rexp(n_customers, lambda_s)
```

Primeiro, rodaram-se os valores no programa desenvolvido neste trabalho, que chamaremos *SimQ*. A seguir o código com os primeiros valores do vetor de saídas **d** dos usuários do sistema.

```
R> simq_output <- simq(arrivals = arrivals,
+    service = service, servers = 2)
R> head(simq_output)
[1] 2.509903 5.041846 5.412492 6.646947 7.185656 9.426389
```

Então, os mesmos valores de **a** e de **s** foram utilizados com o *queuecomputer*, cujo *output*, o vetor de saídas **d**, está representado com suas primeiras observações.

```
R> queuecomputer_output <- queue_step(arrivals = arrivals,
+    service = service, servers = 2)
R> head(sort(depart(queuecomputer_output)))
[1] 2.509903 5.041846 5.412492 6.646947 7.185656 9.426389
```

A saída de ambos os programas resultam em valores exatamente iguais, o que era de se esperar, tendo em vista que se utilizou como base o mesmo algoritmo QDC.

5 Resultados

5.1 Considerações sobre o banco de dados

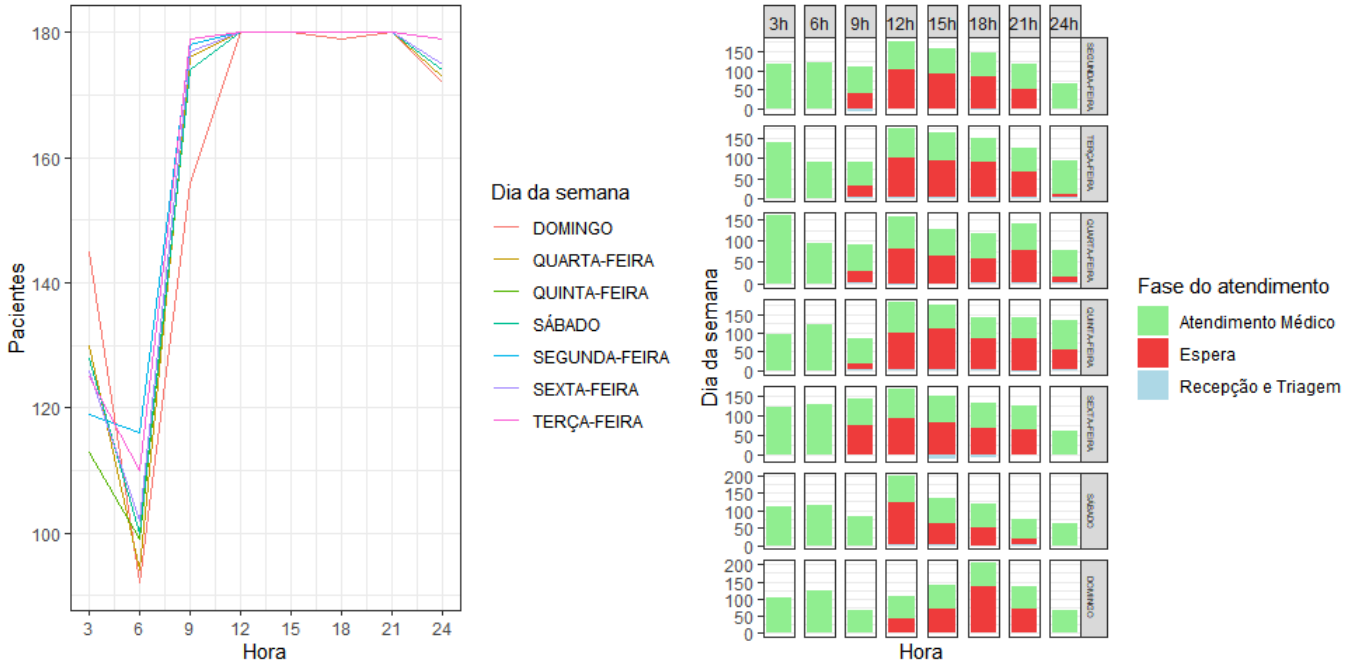
Os dados fornecidos pelo referido hospital são incompletos, tendo em vista que há ocasiões em que médicos não preenchem os horários de atendimento de seus pacientes, ou podem preencher de maneira equivocada, uma vez que não há um sistema automático de compilação dessas informações.

- Não há indicação do tempo de saída do atendimento na recepção;
- Não se sabe se o atendimento na recepção é prioritário (para idosos) ou normal;
- Na triagem, não há nenhuma observação do horário de entrada ou de saída do paciente;
- Não se indica a prioridade de acordo com a urgência médica do caso do paciente;
- Não há registro de qual médico atendeu o paciente;
- Não se sabe se o paciente foi para o laboratório de exames ou uma sala de medicação e quanto tempo passou lá, e, se for o caso, se retornou a um médico.

Além disso, pela ausência de detalhamento dos horários de atendimento dos pacientes, os dados referentes ao tempo de espera na análise descritiva (Figura 6) são o somatório total do tempo de espera do paciente em todo o processo do atendimento na emergência, ou seja, somam-se o tempo de espera para atendimento na recepção, na triagem e espera até que finalmente o paciente seja recebido nos consultórios ou na sala vermelha ou na sala amarela; ainda na mesma figura, o tempo "Atendimento Médico" inclui a consulta, exames e possível retorno ao consultório, uma vez que não há dados específicos referentes a cada uma dessas fases do atendimento médico propriamente dito.

5.2 Análise Descritiva

As figuras a seguir apresentam uma visualização da dinâmica do atendimento de emergência no Hospital das Forças Armadas de Brasília ao longo do tempo demarcado de três em três horas, de forma que possam ser avaliadas alternativas de alocação da força de trabalho do hospital em função do turno (diurno/noturno), ou dia da semana.



(a) (b)

Figura 6: (a) Quantidade média de pacientes que dão entrada na emergência do HFA por dia e horário. (b) Tempo médio em cada fase de atendimento da emergência do HFA por dia e horário.

Pelo limite abrupto observado para todos os dias da semana em 180 pacientes na Figura 6(a), há uma indicação de equívoco no banco de dados. Fora isso, pode-se constatar que o horário com menor quantidade de pacientes que chegam à emergência do hospital é por volta das seis da manhã, em todos os dias da semana, tendo em vista que cada linha segue uma trajetória similar.

O gráfico da Figura 6(b) indica que não há espera até as seis horas da manhã e que o maior tempo de espera ocorre por volta do meio dia, que deve ser o "horário de pico" do hospital, exceto nos domingos, em que a maior espera ocorre ao redor das 18h. O tempo na recepção não representa parte significativa do tempo total no sistema.

5.3 Funcionamento do Modelo Computacional

O modelo computacional tem como *input* os valores gerados aleatoriamente descritos na Seção 4.1 e a quantidade escolhida de funcionários em cada estágio do atendimento em cada turno e retorna *dataframes* com os tempos e informações de cada paciente, como tempo de chegada naquele estágio, tempo em atendimento (tempo de serviço) e a prioridade.

	chegadas	recepcao	atendimento	saidas	enfermeiro	prioridade
30	472.53	472.53	0.76	473.29	1	Normal
31	472.85	472.85	1.07	473.92	2	Normal
32	472.88	473.29	1.13	474.43	1	Normal
33	476.85	476.85	1.79	478.64	2	Normal
34	482.45	482.45	1.32	483.77	1	Normal
35	482.74	482.74	1.58	484.32	2	Normal

Tabela 5: Seis observações de uma tabela referente ao atendimento na recepção gerada pelo modelo.

	chegadas	consultorio	atendimento	saidas	médico	prioridade
30	506.98	513.55	10.25	523.81	4	Não urgente
31	508.31	509.24	13.79	523.03	2	Pouco urgente
32	509.78	521.73	18.16	539.89	3	Não urgente
33	519.82	523.81	21.41	545.21	4	Não urgente
34	522.11	523.03	15.28	538.31	2	Pouco urgente
35	529.75	529.75	14.88	544.63	1	Não urgente

Tabela 6: Seis observações de uma tabela referente ao atendimento nos consultórios médicos gerada pelo modelo.

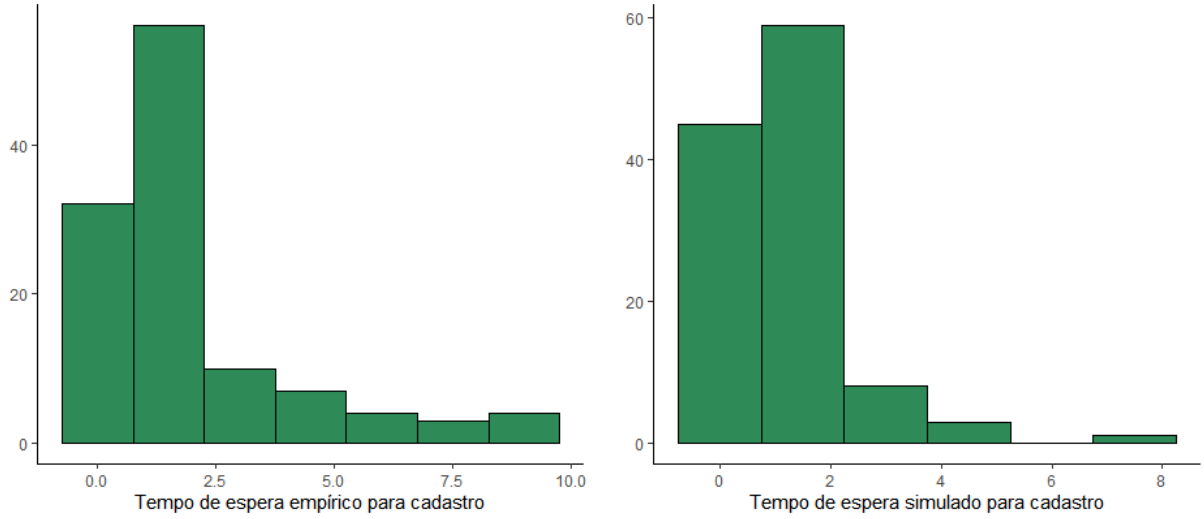
A partir dessas tabelas, são calculadas as medidas de desempenho W, o tempo médio no sistema (em minutos), W_q, o tempo médio de espera na fila, L, a quantidade média de pessoas no sistema, e L_q, o comprimento médio da fila.

Devido à falta de dados disponíveis sobre o caminho detalhado dos pacientes dentro da emergência - ou seja, em que horário determinado paciente entrou e saiu em cada estágio - recepção, triagem, consulta médica, etc. - não se pode testar a adequabilidade das distribuições utilizadas para descrever o sistema nem avaliar a qualidade da estimação do modelo computacional implementado, a não ser por uma comparação com o pacote já estabelecido do R *queuecomputer*, o que é discutido na Seção 4.4.1.

Apesar disso, as medidas descritivas e os histogramas a seguir mostram de uma maneira geral o ajuste aos dados fornecidos pelo hospital pelo menos em relação ao atendimento na recepção, em que há dados completos ao menos para o horário de chegada e horário de início do atendimento.

	Medidas					
Tempo de espera	Min	Q_1	Md	μ	Q_3	Max
Empírico	0	0	1	2.19	3	9
Simulado	0.08	0.55	1.02	1.22	1.54	7.96

Tabela 7: *Summary* do tempo de espera empírico e do temo de espera simulado.



(a) (b)

Figura 7: (a) Histograma do tempo de espera empírico para ser atendido na recepção. (b) Histograma do tempo de espera simulado para ser atendido na recepção.

5.4 Estudo para remanejamento mais eficiente da força de trabalho do hospital

Devido à precariedade do conjunto de dados, não há como avaliar o ajuste do modelo computacional desenvolvido a dados reais nem a escolha das distribuições e parâmetros, feita com base em uma média de tempo para a recepção, e uma média do tempo total de atendimento médico. Sendo assim, esta seção tem o objetivo apenas de avaliar aspectos metodológicos da ferramenta implementada, não podendo servir como instrumento para decisões administrativas do hospital. Quando os dados completos e corrigidos do hospital estiverem disponíveis, a ferramenta poderá ser utilizada com poucas modificações.

	Estrutura original		Alteração 1		Alteração 2		Alteração 3	
Local	D	N	D	N	D	N	D	N
Recepção	2	1	2	1	2	1	2	1
Triagem	1	1	1	1	1	1	1	1
Consultórios	3	3	4	4	2	4	4	2
Sala Vermelha	2	2	2	2	2	2	2	2
Sala Amarela	2	2	1	1	2	2	2	2

Tabela 8: Estrutura da força de trabalho do atendimento de emergência do hospital das forças armadas, contendo a estrutura original em vigência no funcionamento do hospital e três alterações simuladas, sendo D, o turno diurno e N, o noturno.

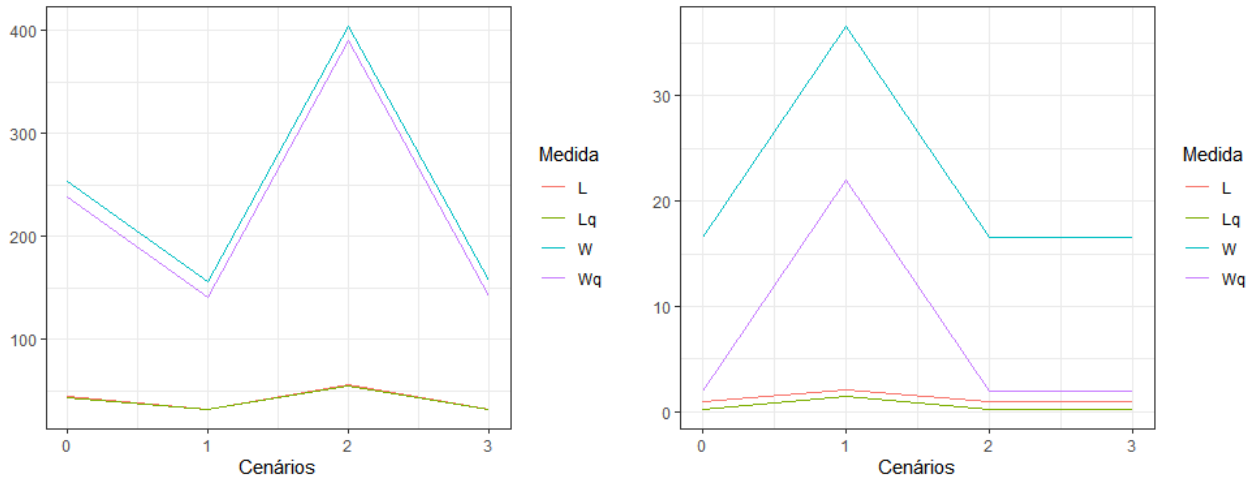
Como o tempo de atendimento na recepção e na triagem não representam parte significativa do tempo total no sistema (apenas 5.15% desse tempo, conforme a Tabela 9), decidiu-se estudar alterações apenas na quantidade de médicos, ou seja, verificar mudanças na quantidade de médicos por turno nos consultórios e nas salas amarela e vermelha.

Recorde que W é o tempo médio no sistema (em minutos), W_q é o tempo médio de espera na fila, L é a quantidade média de pessoas no sistema, e L_q é o comprimento médio da fila.

	Medida			
Fase de Atendimento	W	W_q	L	L_q
Recepção	6.8	5.5	3.07	2.34
Triagem	8.8	7.5	3.42	2.74
Consultório	253.15	238.22	44.08	43.82
Sala Vermelha	17.43	4.25	1.3	0.47
Sala Amarela	16.54	1.92	1.02	0.2

Tabela 9: Medidas de desempenho do atendimento em cada estágio retornadas pelo modelo de acordo com a estrutura atual de atendimento de emergência do HFA de Brasília.

Verificamos, pela Figura 8, que, com a mesma quantidade de médicos, pode ser atingida uma eficiência consideravelmente melhor apenas alocando um médico do turno noturno para o trabalho no turno diurno em cada consultório, resultando em uma diminuição do tempo de espera nos consultórios médicos de aproximadamente 40.3%.



(a) (b)

Figura 8: (a) Desempenho do atendimento nos consultórios médicos de acordo com os diferentes cenários simulados, com 0 representando a estrutura original de atendimento e 1,2 e 3 representando as alterações estudadas na estrutura de atendimento. (b) Desempenho do atendimento na sala amarela de acordo com os diferentes cenários simulados, de maneira análoga à 8(a).

Além disso, caso seja possível aumentar o tempo de espera na sala amarela para 22 minutos, é mais uma opção alocar um médico em serviço nessa sala para atender nos consultórios a fim de aumentar a eficiência nos consultórios médicos, como observamos pelos gráficos da Figura 8. Porém, não se sabe se é possível retirar um médico das salas de atendimento mais urgente, pois pode haver alguma regra que imponha uma quantidade mínima de profissionais em cada sala.

De qualquer forma, o simples remanejamento de um médico do turno noturno para o diurno pode ser uma boa opção para a melhora do atendimento de emergência do hospital, sem que prejudique o atendimento na sala amarela e resultando em um aumento quase igual de eficiência entre os cenários 1 e 3.

5.5 Aplicativo *Dashboard*

O aplicativo serviria como ferramenta de gestão para o hospital, tanto para a avaliação do atendimento na emergência, como para verificar alterações na estrutura de atendimento que pudessem tornar o processo mais eficiente, com menor quantidade de funcionários, menores filas e menor tempo de espera.

Feito com o pacote do R para R Markdown *flexdashboard*, funciona de maneira muito simples: basta escolher os parâmetros de quantidade de funcionários em cada fase de atendimento (recepção, triagem, etc.) e em cada turno e o painel relaciona medidas de desempenho e gráficos úteis para a visualização do atendimento.

É dividido em três seções: a primeira é um painel com medidas gerais do sistema (Figura 9), a segunda contém as medidas do atendimento anterior à consulta médica, ou seja, as medidas da recepção e da triagem, e a terceira são medidas dos consultórios e das salas amarela e vermelha (Figura 10).

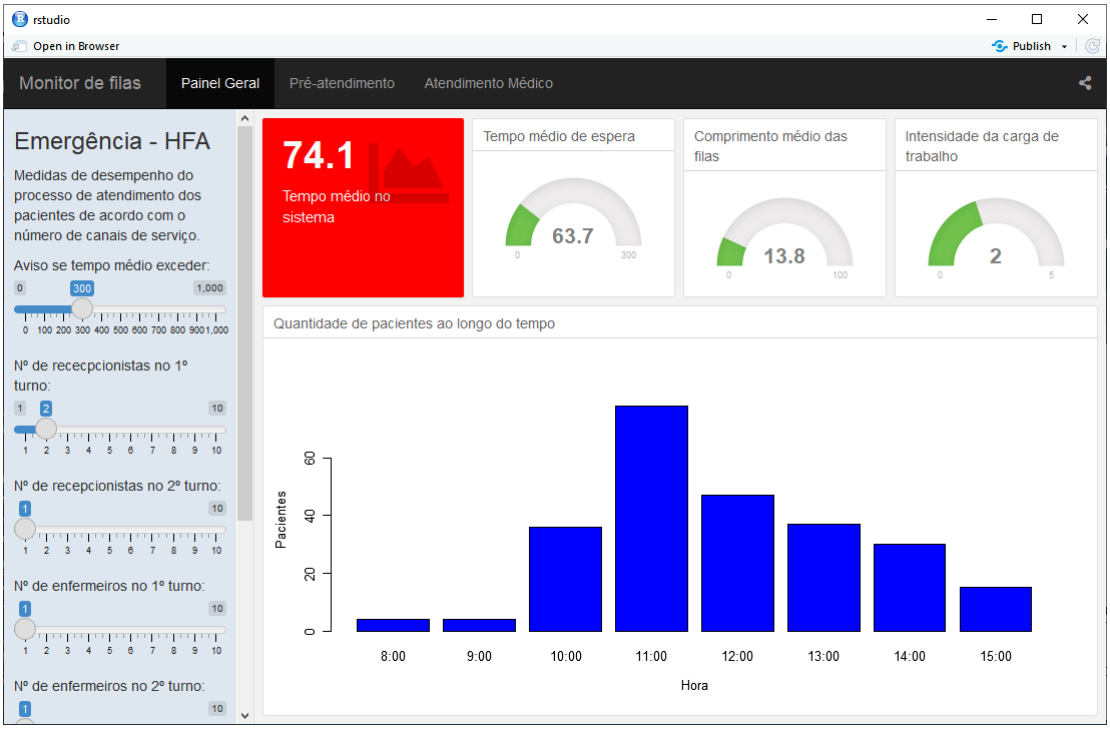


Figura 9: *Dashboard* de medidas de desempenho gerais da emergência do HFA.

Embora não seja o escopo desta pesquisa, o aplicativo ainda pode ser estendido para realizar um monitoramento em tempo real da fila, oferecendo uma excelente ferramenta não apenas para o hospital, mas também para a população em geral que necessite de atendimento médico e poderia verificar se o hospital está com fila de espera ou tempo de espera muito grande, dentre outras informações. Se um sistema de monitoramento como esse fosse implementado em todos os hospitais da cidade, seria fácil alocar os pacientes nas instituições da forma mais eficiente possível.

A medida de intensidade da carga de trabalho, também chamada de intensidade de tráfico, refere-se à quantidade média de pessoas em atendimento, isto é, se essa medida resultar em valor 2, por exemplo, para a triagem, isso quer dizer que, em média, duas pessoas estão sendo atendidas na triagem a todo momento. Para saber a intensidade de carga de trabalho para cada servidor, basta dividir pela quantidade de servidores em atendimento.

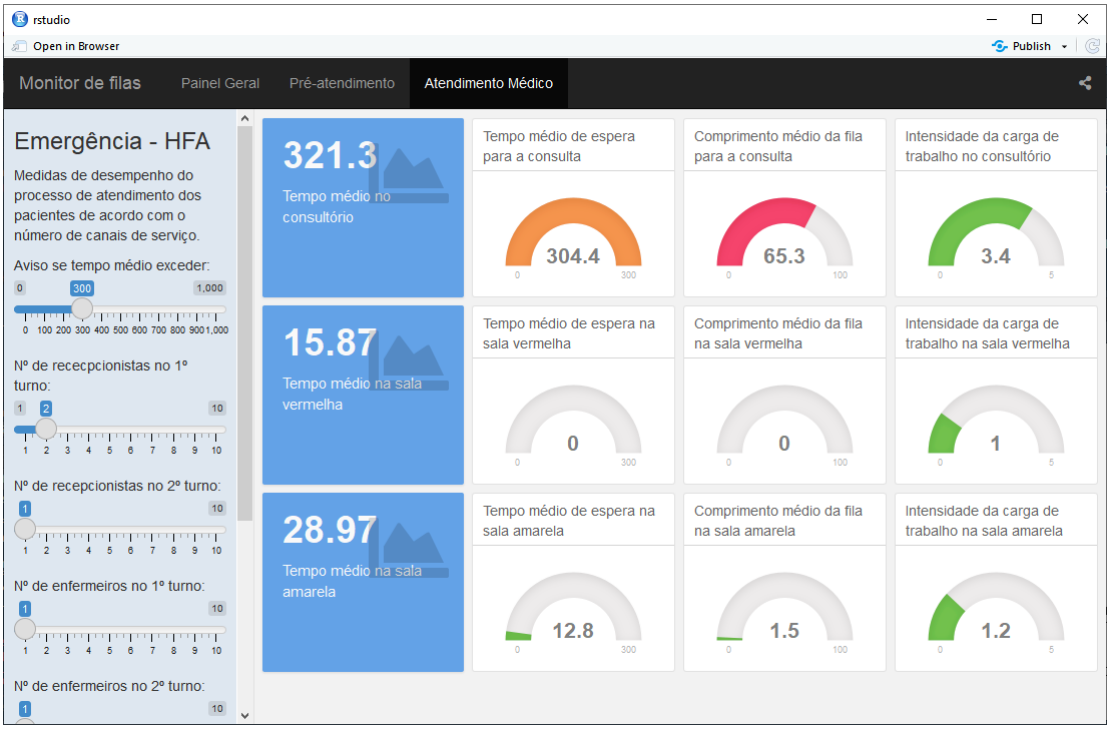


Figura 10: *Dashboard* de medidas de desempenho do atendimento médico de emergência do HFA.

5.6 Animação do caminho percorrido pelos pacientes

Para a visualização do fluxo de pacientes que são atendidos pela emergência do hospital ao longo do dia, foi feita uma animação, utilizando o pacote *gganimate* do R, em que cada fase de atendimento estudada (recepção, triagem e atendimento médico nas respectivas salas ou consultórios) é representada por um bloco dentro de um conjunto de coordenadas cartesianas.

Os pacientes recebem coordenadas (x, y) de acordo com o local em que foram atendidos: na sala de espera, recebem valores aleatórios de coordenadas limitados dentro do bloco sala de espera. Ou seja, para sua posição em espera para ser atendido na recepção, o paciente recebe aletaoriamente um vetor bidimensional

$$(x, y | min(\mathbf{x}_{espera}) < x < max(\mathbf{x}_{espera}) \quad \& \quad min(\mathbf{y}_{espera}) < y < max(\mathbf{y}_{espera})).$$

As coordenadas para cada um dos locais de atendimento indicam o centro do bloco.

Para representar o caminho percorrido entre os locais, uma função traça um segmento de reta que liga um bloco a outro, formado pela passagem do tempo do sistema. O tempo em que o paciente ficou em cada um desses locais é proporcional ao tempo estimado pelo modelo.

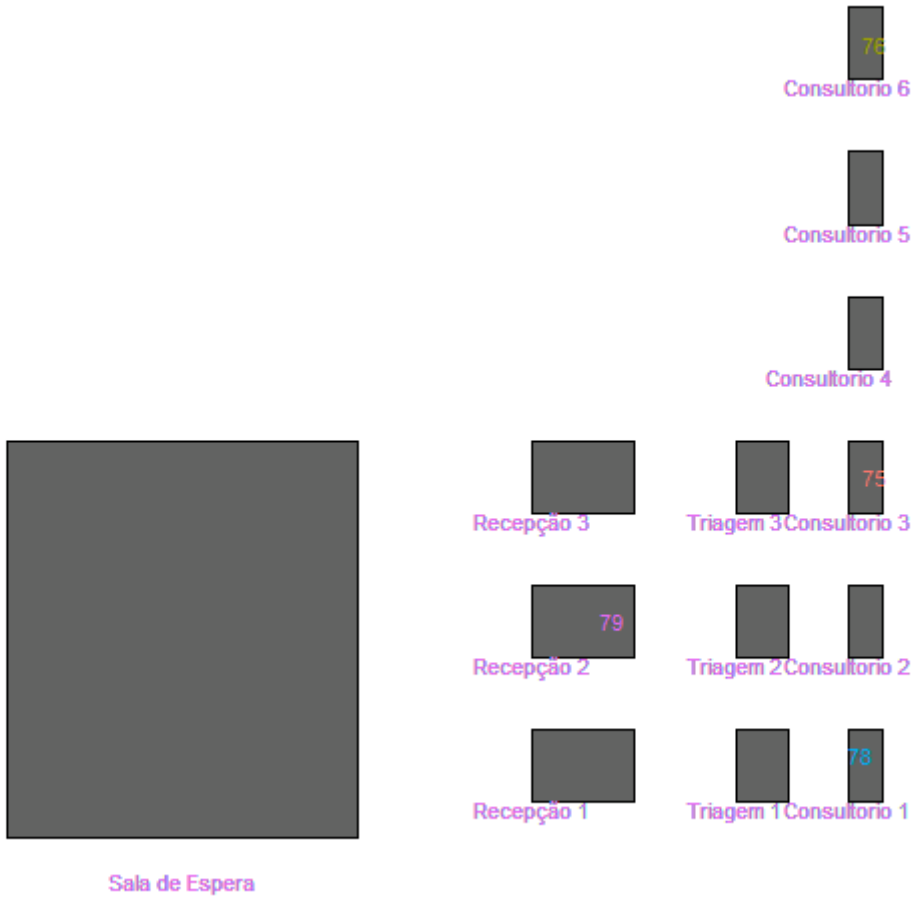


Figura 11: Momento da animação que mostra o caminho percorrido pelos pacientes simulados de número 75 a 79. Observe que o paciente 77 já terminou seu atendimento enquanto os pacientes 75, 76 e 78 estão ainda sendo atendidos nos consultórios e o paciente 79 ainda está sendo atendido na recepção.

Caso o leitor queira visualizar este trecho da simulação do caminho dos pacientes, pode acessar o link:

`https://drive.google.com/drive/folders/1rS0KxvgycdR42JIJDJJAKVJTNV4YvR6V?usp=sh`

.

6 Conclusão

A revisão teórica acerca de cadeias de filas foi importante a fim de aprofundar conceitos sobre o tema, como o estudo dos modelos avançados e, de forma direta, foram utilizados conhecimentos importantes acerca de simulação e estimação de processos. De forma indireta, o estudo da Teoria de Filas consolidou o entendimento das limitações intrínsecas às aplicações dos modelos analíticos desenvolvidos; mesmo os de aspecto mais geral, que comportam quaisquer distribuições, não são capazes de descrever sistemas mais complexos do que um atendimento linear independente, ou seja, não são capazes de abordar processos em que haja interdependência entre diferentes estágios de atendimento dentro do mesmo sistema. Dessa forma, foi validado o esforço de descrever computacionalmente o sistema de atendimento adotado no hospital.

A aplicação prática deixou a desejar, em parte, devido à baixa qualidade do banco de dados construído pelo hospital a fim de monitorar o atendimento no serviço que presta na área de emergência. Como apontado anteriormente, há dois problemas principais com os dados do hospital:

- Erros de registro: quando o dado registrado não está de acordo com o que de fato ocorreu. Há dados em que, supostamente, o paciente esperou por dias o atendimento médico. Isso se deve, em parte, pela necessidade de um médico ou técnico registrar manualmente a informação no sistema.
- Ausência de dados básicos sobre o processo: não se tem registro dos horários de chegada e de saída de atividades intermediárias como sala de medicação, sala de exames, triagem e o fim do atendimento na recepção. Também não se tem registro da estrutura de funcionários vigente em cada horário, nem qual médico atendeu o paciente.

Constata-se uma grande discrepância entre o fluxo apresentado na Figura 5 e os dados coletados pelo sistema em uso. Os processos estão devidamente mapeados, mas não se tem dados empíricos sobre o funcionamento do sistema. Cada movimentação do paciente precisa ser registrada, de modo que se saiba, com precisão, todo o histórico do atendimento.

Dessa forma, sugerimos ao hospital as seguintes recomendações: cadastro do atendimento de cada médico; preenchimento automático do horário das consultas; preenchimento dos horários de atendimento na triagem; preenchimento dos horários de atendimento em salas de exame; preenchimento dos horários de atendimento em salas de medicação; indicação de retorno ao consultório de paciente previamente atendido; e indicação da quantidade de médicos atendendo em cada especialidade.

A partir da coleta de dados completos e de qualidade, o hospital terá seu trabalho

de gestão facilitado, tanto na avaliação do serviço prestado como também para avaliar alternativas na estrutura de atendimento que tornem o serviço mais eficiente, benéfico para o próprio hospital, que terá um atendimento mais organizado e sinérgico, e para a população que necessita desse serviço de saúde. A automação do processo de coleta desses dados é de suma importância, uma vez que possibilita a análise e monitoramento do atendimento e dispensa os médicos de terem que ocupar seu tempo com atividades administrativas.

Apesar de não poder ser aproveitada a aplicação prática do modelo desenvolvido, a sua própria construção foi de imensa valia para o aprimoramento do conhecimento acerca de programação e modelagem. O programa desenvolvido replicou satisfatoriamente resultados obtidos com outros programas que serviram de *benchmark*, com a implementação das funcionalidades de se poder alterar a quantidade de funcionários/servidores em função do tempo e de poder descrever atendimentos que levem em conta a classificação de prioridade na fila.

Além do modelo em si, o desenvolvimento do *dashboard* e da animação demonstram seu potencial de aplicação em situações reais, em empresas, consultórios, hospitais ou qualquer estabelecimento que tenha como aspecto importante um processo de espera para seus clientes.

6.1 Sugestões para trabalhos futuros

- Implementação de nova funcionalidade que permita o retorno de um usuário a uma fase anterior de atendimento;
- Estudo mais aprofundado da estimação dos parâmetros e das distribuições que descrevem o processo;
- Funcionamento do *Dashboard shiny* em tempo real.

Referências

BARLOW, R. E.; PROSCHAN, F. Statistical theory of reliability and life testing.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2001. v. 2.

CASELLA, G.; BERGER, R. L. Inferência estatística. *São Paulo: Cengage Learning*, 2010.

EBERT, A. et al. Computationally efficient simulation of queues: The r package queuecomputer. 2019.

FOX, B. L. Fitting 'standard' distributions to data is necessarily good: Dogma or myth?

GROSS, D.; HARRIS, C. M. *Fundamentals of Queueing Theory*. [S.l.: s.n.], 1997.

GROSS, D.; JUTTIJUDATA, M. Sensitivity of output measures to input distributions in queueing simulation modeling.

JUTTIJUDATA, M. Sensitivity of output performance measures to input distributions in queueing simulation modeling.

KELTON, W. D. Input data collection and analysis.

KENDALL, M. The analysis of economic time series, part i: Prices. *Journal of the Royal Statistical Society*.

LAW, A. M.; KELTON, W. D. Simulation modeling and analysis.

LEEMIS, L. M. Discrete-event simulation input process modeling.

LINDLEY D, V. The theory of queues with a single server. *Proc. Camb. Phil. Soc.*

LITTLE, J. D. C. A proof for the queuing formula: $L = \lambda W$. 1961.

MARCHAL, W. G. Some simpler baound on the mean queueing time. *Oper. Res.*

MARSHALL, K. T. Some inequalitites in queueing. *Oper. Res.*

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: ⟨http://www.R-project.org/⟩.

RIZZO, M. L. *Statistical computing with R*. [S.l.]: Chapman and Hall/CRC, 2007.

ROBERT, C. P.; CASELLA, G.; CASELLA, G. *Introducing monte carlo methods with r*. [S.l.]: Springer, 2010. v. 18.

SMITH, W. L. On the distribution of queueing times. *Proc. Camb. Phil. Soc.*

SMITH, W. L.; WILKINSON, W. E. Proceedings symposium on congestion theory.

7 Apêndices

7.1 Código para simulação do problema da seção 3.4.5 - pacote *Simmer*

```
set.seed(223)
t0 <- trajectory("my trajectory") %>%
  ## add an intake activity
  seize("recepção", 1) %>%
  timeout(function() rnorm(1, 3)) %>%
  release("recepção", 1) %>%
  ## add a consultation activity
  seize("enfermagem", 1) %>%
  timeout(function() rnorm(1, 10, 2)) %>%
  release("enfermagem", 1) %>%
  ## add a planning activity
  seize("médico", 1) %>%
  timeout(function() rnorm(1, 15, 4)) %>%
  release("médico", 1)

env <- simmer("SuperDuperSim") %>%
  add_resource("recepção", 3) %>%
  add_resource("enfermagem", 4) %>%
  add_resource("médico", 5) %>%
  add_generator("paciente", t0, function(){rnorm(1, 3)}) %>%
  add_generator("idoso", t0, at(sample(c(1:100),5)), priority=1) %>%
  run(until=200)

resources <- get_mon_resources(env)
arrivals <- get_mon_arrivals(env)

plot(resources, metric="usage", c("recepção","enfermagem","médico"),
     items = c("server","queue"),
     steps = TRUE)
plot(resources, metric="utilization", c("recepção","enfermagem","médico"))
plot(arrivals, metric="waiting_time")
```

```
mean(arrivals$end_time - arrivals$start_time)
mean(arrivals$end_time - arrivals$start_time - arrivals$activity_time)

set.seed(223)
t0 <- trajectory("my trajectory") %>%
  ## add an intake activity
  seize("recepção", 1) %>%
  timeout(function() rnorm(1, 3)) %>%
  release("recepção", 1) %>%
  ## add a consultation activity
  seize("enfermagem", 1) %>%
  timeout(function() rnorm(1, 10, 2)) %>%
  release("enfermagem", 1) %>%
  ## add a planning activity
  seize("médico", 1) %>%
  timeout(function() rnorm(1, 15, 4)) %>%
  release("médico", 1)

env <- simmer("SuperDuperSim") %>%
  add_resource("recepção", 1) %>%
  add_resource("enfermagem", 3) %>%
  add_resource("médico", 5) %>%
  add_generator("paciente", t0, function(){rnorm(1, 3)}) %>%
  add_generator("idoso", t0, at(sample(c(1:100),5)), priority=1) %>%
  run(until=200)

resources <- get_mon_resources(env)
arrivals <- get_mon_arrivals(env)

plot(resources, metric="usage", c("recepção","enfermagem","médico"),
     items = c("server","queue"),
     steps = TRUE)
plot(resources, metric="utilization", c("recepção","enfermagem","médico"))
plot(arrivals, metric="waiting_time")


mean(arrivals$end_time - arrivals$start_time)
mean(arrivals$end_time - arrivals$start_time - arrivals$activity_time)
```

7.2 Descrição das variáveis do conjunto de dados do hospital HFA

Variável	Descrição	Categorias
CD_ATENDIMENTO	Cadastro do atendimento	-
CD_PACIENTE	Cadastro do paciente	-
DH_RETIRADA_SENHA	Data e horário em que o paciente retirou a senha	-
DH_CADASTRO	Data e horário em que o paciente foi cadastrado na recepção	- -
DH_INI_CLASSIFICACAO	Data e horário em que o paciente foi atendido na triagem	-
DH_FIM_CLASSIFICACAO	Data e horário em que o paciente terminou de ser atendido na triagem	-
DH_INI_ATEND_MED	Data e horário em que o paciente foi atendido pelo médico(a)	-
DH_FIM_ATEND_MED	Data e horário em que o paciente terminou de ser atendido pelo médico(a)	-
DH_ENTRADA_SL_AMARELA	Data e horário em que foi dada a entrada do paciente na sala amarela	-
DH_SAIDA_SL_AMARELA	Data e horário em que o paciente terminou de ser atendido na sala amarela	-
DH_ENTRADA_SL_VERMELHA	Data e horário em que o paciente foi atendido na sala vermelha	-
DH_SAIDA_SL_VERMELHA	Data e horário em que o paciente terminou de ser atendido na sala vermelha	-
DH_ALTA_MEDICA	Data e horário em que o paciente terminou de ser atendido na sala vermelha	-
DH_ADMISSAO	Data e horário em que o paciente foi admitido para internação	-
ESPECIALIDADE	Especialidade médica a que o paciente foi transferido	RADIOLOGIA ORTOPEDIA/TRAUMATOLOGIA CLINICA GERAL CIRURGIA GERAL GINECOLOGIA OFTALMOLOGIA GINECOLOGIA/OBSTETRICIA CLINICA MEDICA NEUROCIRURGIA ODONTOLOGIA ORTODONTIA UROLOGIA NEUROLOGIA ENFERMAGEM NUTRICIONISTA CARDIOLOGIA GASTROENTEROLOGIA OTORRINOLARINGOLOGIA CIRURGIA B.M.F. (ODONTO) CIRURGIA VASCULAR COLOPROCTOLOGIA PSIQUIATRIA PNEUMOLOGIA/TISIOLOGIA CIRURGIA PLASTICA TECNICO DE ENFERMAGEM OBSTETRICIA FISIATRIA CIRURGIA TORACICA IMOBILIZACAO ORTOPEDICA ANESTESIOLOGIA
DIA_DA_SEMANA	Dia da semana em que o paciente deu entrada na emergência do hospital	SEGUNDA-FEIRA TERÇA-FEIRA QUARTA-FEIRA QUINTA-FEIRA SEXTA-FEIRA SÁBADO DOMINGO

8 Anexos

8.1 Demonstração - Função de distribuição acumulada do tempo de espera na fila ($W_q(t)$) para o modelo $M/M/c/k$

Seja $W_q(t)$ a função de distribuição acumulada de T_q, onde T_q é a variável aleatória contínua que representa o tempo de espera na fila de um usuário qualquer. Então, $W_q(t)$ representa a probabilidade de um usuário qualquer aguardar na fila por um tempo máximo de $t \geq 0$.

$$W_q(t) = P(T_q \leq t) = W_q(0) + \sum_{n=c}^{K-1} P\{\text{n-c+1 atendimentos em } \leq t| \\ \text{chegada n no sistema}\}q_n,$$

q_n é a probabilidade de haver n no sistema dado que uma chegada está prestes a ocorrer, i.e.,

$$\begin{aligned} q_n &= P(n \text{ no sistema}|\text{chegada prestes a ocorrer}) \\ &= \frac{P(\text{chegada prestes a ocorrer}|n \text{ no sistema})p_n}{\sum_{n=0}^{K} P(\text{chegada prestes a ocorrer}|n \text{ no sistema})p_n} \\ &= \lim_{\Delta t \to \infty} \left\{ \frac{[\lambda \Delta t + o(\Delta t)]p_n}{\sum_{n=0}^{K-1}[\lambda \Delta t + o(\Delta t)]p_n} \right\} \\ &= \lim_{\Delta t \to \infty} \left\{ \frac{[\lambda + o(\Delta t)/\Delta t]p_n}{\sum_{n=0}^{K-1}[\lambda + o(\Delta t)/\Delta t]p_n} \right\} \\ &= \frac{\lambda p_n}{\lambda \sum_{n=0}^{K-1} p_n} \\ &= \frac{p_n}{1 - p_k} \qquad (n \leq K-1) \end{aligned}$$

em que Δt é um incremento de tempo e $o(t)$ é a probabilidade de que ocorra mais de uma chegada entre t e Δt, quantidade que se torna negligenciável quando $\Delta t \to \infty$, i.e.,

$$\lim_{\Delta t \to \infty} o(\Delta t)/\Delta t = 0$$

Assim,

$$W_q(t) = W_q(0) + \sum_{n=c}^{K-1} q_n \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x} dx$$
$$= W_q(0) + \sum_{n=c}^{K-1} q_n \left(1 - \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x} dx\right)$$

Pode-se demontrar que (Gross e Harris - *Queueing Thoery, section 1.7*),

$$\int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x} dx = \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}$$

Finalmente,

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

8.2 Outros modelos básicos

M/M/1/k

$$P_0 = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}, \text{ se } \rho \neq 1 \\ \frac{1}{K+1}, \text{se } \rho = 1. \end{cases}$$

$$P_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, \text{ se } \rho \neq 1 \\ \frac{1}{K+1}, \text{se } \rho = 1. \end{cases}$$

As medidas de desempenho para este sistema, também derivadas de M/M/c/K, estão dadas a seguir.

Número médio de usuários na fila (L_q)

$$L_q = \begin{cases} \frac{(\rho)}{1-\rho} - \frac{\rho(K\rho^K+1)}{1-\rho^{K+1}}, \text{ se } \rho \neq 1 \\ \frac{K(K-1)}{2(K+1)}, \text{se } \rho = 1. \end{cases}$$

Número médio de usuários no sistema(L)

$$L = L_q + (1 - P_0)$$

Note que essa relação implica que $1 - P_0 = \lambda(1 - P_K)/\mu$, que pode ser reescrita como $\mu(1 - P_0) = \lambda(1 - P_K)$, verificando que a taxa de saída efetiva do sistema é igual à taxa efetiva de chegada.

Tempo médio de permanência no sistema (W)

Para usarmos a fórmula de Little, temos que considerar a limitação da capacidade do sistema. Rejeições ocorrem a uma taxa λP_k cada vez que o sistema atinge o estado k. Assim sendo, a taxa de ingressos $\lambda^{'}$ não coincide com a taxa de chegadas λ,

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

que, substituída na respectiva fórmula de Little, nos dá,

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

Tempo Médio de espera na fila (W_q)

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

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

A demonstração para a função de distribuição acumulada do tempo de espera na fila para este modelo encontra-se mais a frente no apêndice.

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

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

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

Caso particular M/M/1/1/FIFO

Neste caso, não há processo de espera, pois o sistema não admite fila, ou seja, só há um usuário no sistema sendo atendido até que ele saia e outro usuário possa ingressar no sistema. Existem, assim, apenas os estados $n = 0$ e $n = 1$. Para qualquer ρ,

$$P_1 = \rho P_0.$$

Sendo $P_0 + P_1 = 1$,

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

e

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

Demonstração - Função de distribuição acumulada do tempo de espera na fila ($W_q(t)$) para o modelo $M/M/1/k$

Seja N^* a variável que representa o número de usuários que efetivamente ingressam no sistema. Sua distribuição, q_n, é a distibuição da v.a. N definida no modelo anterior, truncada a direita em n=k,

$$q_n = \begin{cases} \frac{P_n}{1-P_k}, 0 \leq n \leq k-1 \\ 0, n \geq k. \end{cases}$$

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

"n serviços completados até t"equivale a "o tempo para completar n serviços é menor do que t".

Se $S_{(n)}$ é a variável aleatória contínua que representa a soma dos tempos de atendimento de n usuários consecutivos, sendo os tempos de serviço independentes e exponencialmente distribuídos à taxa μ, então $S_{(n)}$ segue uma distribuição de Erlang de parâmetros n e μ, e

$$
\begin{aligned}
W_q(t) &= W_q(0) + \sum_{n=1}^{k-1} q_n \int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx \\
&= W_q(0) + \sum_{n=1}^{k-1} q_n \left[1 - \int_0^\infty \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx \right] \\
&= W_q(0) + \sum_{n=0}^{k-2} q_{n+1} \left[1 - \int_0^\infty \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx \right] \\
&= W_q(0) + \sum_{n=0}^{k-2} q_{n+1} \left[1 - \sum i = 0^n \frac{(\mu t)^i}{(i)!} e^{-\mu t} \right] \\
&= 1 - \sum_{n=0}^{k-2} q_{n+1} \sum_{i=0}^{n} \frac{(\mu t)^i}{(i)!} e^{-\mu t},
\end{aligned}
$$

pois

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

e $W_q = q_0$.

M/M/c/∞

Pelo fato de que as chegadas e os atendimentos neste modelo são Processos de Nascimento e Morte, as taxas de chegadas e de atendimento são respectivamente dadas por:

Taxa de Ingresso:

$$\lambda_n = \lambda,\ n \geqslant 0$$

Taxa de Atendimento:

$$\mu_n = \begin{cases} n\mu, \text{ se } 1 \leqslant n < c \\ c\mu, \text{ se } n \geqslant c. \end{cases}$$

Probabilidade no Regime Estacionário:

$$P_n = \begin{cases} P_0 \frac{r^n}{n!}, \text{ se } 1 \leqslant n < c \\ P_0 \frac{r^n}{c^{n-c}c!}, \text{ se } n \geqslant c. \end{cases}$$

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{\infty} \frac{\lambda^n}{c^{n-c}c!\mu^n} \right)$$

Número Médio de Usuários na Fila:

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

Número Médio de Usuários no sistema:

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

Tempo Médio de Espera na Fila:

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

Tempo Médio de Espera no sistema:

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

Função de distribuição acumulada do tempo de espera na fila

De maneira análoga ao modelo M/M/c/K, sendo que $q_n = p_n$, pois não há clientes rejeitados p_K, tem-se

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

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

Diretamente de $W_q(t)$,

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

M/M/1

As chegadas e os atendimentos expressam um processo de nascimento e morte, sendo que somente um único evento pode acontecer em períodos de tempo pequenos. As taxas de chegada e de atendimento são constantes, dadas, respectivamente, por

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

e

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

Em qualquer processo Markoviano que se encontra em estágio estacionário, a probabilidade do sistema estar no estado n no instante t (ou seja, o número de usuários ser n no instante t) é

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

Pela distribuição estacionária de processos de nascimento e morte, temos,

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

e

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

Como temos uma série geométrica, se $\frac{\lambda}{\mu} < 1$,

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

Seja $\rho = \frac{\lambda}{\mu}$ a taxa de ocupação/utilização do sistema, que, ao ser substituída nas fórmulas acima, resulta em

$$P_n = \rho^n (1-\rho)^n \,\forall\, n \geq 0.$$

Assim, vemos que o número de usuários do sistema segue uma distribuição geométrica (com a probabilidade de n "insucessos"antes do primeiro "sucesso") de parâmetro $(1-\rho)$, que possui o valor esperado de $\frac{\rho}{1-\rho}$.

Número médio de usuários no sistema(L)

Seja N a variável aleaória discreta que representa o número de usuários no sistema no regime de estacionariedade, com distribuição de probabilidade P_n, $n \geq 0$ e valor esperado L. Então,

$$L = E(N) = \sum_{n=0}^{\infty} nP_n$$

Usando a equação com a taxa de ocupação/utilização, temos

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

Se $\rho < 1$,

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

Portanto,

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

o que atesta o valor esperado da distribuição geométrica $P_n = \rho^n(1-\rho)^n$.

Número médio de usuários na fila (L_q)

Seja N_q a variável aleaória discreta que representa o número de usuários na fila no regime de estacionariedade, com valor esperado L_q. Então,

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

seguindo,

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

Probabilidade de se ter mais do que k elementos no sistema

Serve para avaliar a necessidade de incluir estrutura física para que as pessoas possam esperar, como cadeiras, banheiros, bebedouros...

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

portanto,

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

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

Seja $W_q(t)$ a função de distribuição acumulada de T_q, onde T_q é a variável aleatória contínua que representa o tempo de espera na fila de um usuário qualquer. Então, $W_q(t)$ represent a probabilidade de um usuário qualquer aguardar na fila por um tempo máximo de $t \geq 0$.

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

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

para $t > 0$,

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

"n serviços completados até t"equivale a "o tempo para completar n serviços é menor do que t".

Se $S_{(n)}$ é a variável aleatória contínua que representa a soma dos tempos de atendimento de n usuários consecutivos, sendo os tempos de serviço independentes e exponencialmente distribuídos à taxa μ, então $S_{(n)}$ segue uma distribuição de Erlang de parâmetros n e μ, e

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

Tendo em vista que $(\mu - \lambda) = \mu(1-\rho)$,

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

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

Derivando a f.d.a., temos

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

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

Analogamente a $W_q(t)$, pode-se obter $W(t)$,

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

Assim, a variável aleatória T, tempo de permanência no sistema, distribui-se exponencialmente com parâmetro $\mu(1-\rho)$ e valor esperado

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

Tempo Médio de espera na fila (W_q)

$$W_q = E(T_q) = \int_0^{\infty} t w_q(t) dt = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{(\mu-\lambda)}.$$

Tempo médio de permanência no sistema (W)

O tempo médio que um usuário qualquer permanece no sistema é a soma do tempo médio de espera na fila com o tempo médio de atendimento, isto é,

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

o que confirma $E(T)$.

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

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

8.3 Demonstrações - modelo $M^{[X]}/M/1$

Se λ_x é a taxa de chegada de um processo de Poisson com "lotes de chegada"de tamanho X, então $c_x = \lambda_x/\lambda$, em que λ é a taxa de chegada composta de todos os lotes, sendo $\lambda = \sum_{i=1}^{\infty} \lambda_i$.

Este processo total, que surge pela sobreposição do conjunto de processos de Poisson com taxas $\{\lambda_x,\, x = 1,\, 2,\, \ldots\}$, é um processo de Poisson múltiplo ou composto.

Assumindo estacionariedade, um conjunto de equações de Chapman-Kolmogorov pode ser derivado para este problema, fazendo surgir as equações balanceadas

$$0 = -(\lambda+\mu)P_n + \mu P_{n+1} + \lambda \sum_{k=1}^{n} P_{n-k} c_k, \qquad (n \geq 1)$$

$$0 = -\lambda P_0 + \mu P_1$$

que são resolvidas utilizando-se funções geratrizes (Gross e Harris - *Queueing Thoery, section 3.1*) ou equações diferenciais quando os tamanhos dos lotes de chegada são pequenos.

Definindo

$$P(z) = \sum_{n=0}^{\infty} P_n z^n \qquad (|z| \leq 1)$$

e

$$C(z) = \sum_{n=0}^{\infty} c_n z^n \qquad (|z| \leq 1)$$

como as funções geratrizes das probabilidades $\{P_n\}$ e das distribuições dos tamanhos dos lotes de chegada $\{c_n\}$, respectivamente.

Multiplicando-se as equações balanceadas pelo z^n apropriado e somando-se as equações resultantes, obtem-se

$$0 = -\lambda \sum_{n=0}^{\infty} P_n z^n - \mu \sum_{n=1}^{\infty} P_n z^n + \frac{\mu}{z} \sum_{n=1}^{\infty} P_n z^n + \lambda \sum_{n=1}^{\infty} \sum_{k=1}^{n} P_{n-k} c_k z^n.$$

Ao fazer uso das propriedades de funções geratrizes, chega-se a $\sum_{n=0}^{\infty} \sum_{k=1}^{n} P_{n-k} c_k z^n = \sum_{k=1}^{n} c_k z^k \sum_{n=k}^{\infty} P_{n-k} z^{n-k} = C(z)P(z)$, e, assim,

$$0 = -\lambda P(z) - \mu[P(z) - P_0] + \frac{\mu}{z}[P(z) - P_0] + \lambda C(z)P(z),$$

que dá

$$P(z) = \frac{\mu P_0(1-z)}{\mu(1-z) - \lambda z[1 - C(z)]} \qquad (|z| \leq 1)$$

Para encontrar P_0 reescreve-se $P(1)$ com $r = \lambda/\mu$

$$P(z) = \frac{P_0}{1 - rz\bar{C}(z)}$$

e notando que $\bar{C}(z) = [1-C(z)]/[1-z]$ é a função geradora das probabilidades complementares dos tamanhos dos lotes de chegada $P(X > x) = 1 - C_x = \bar{C}_x$, desde que $1/(1-z)$ é a função geradora de 1 e $C(z)/(1-z)$ é a função geradora das probabilidades cumulativas C_x, o que pode ser visto notando-se que $\sum_{x=1}^{\infty} C_x z^x = \sum_{x=1}^{\infty} \left(\sum_{i=1}^{x} c_i\right) z^x = \left(\sum_{i=1}^{\infty} c_i z^i \frac{1}{1-z}\right)$.

Aplicando-se a regra de l'Hôpital uma vez, tem-se $\bar{C}(1) = E[X]$, e

$$1 = P(1) = \frac{P_0}{1 - r\bar{C}(1)}$$

que, finalmente, dá

$$P_0 = 1 - rE[X] = 1 - \rho$$

Aplicando-se l'Hôpital novamente, encontra-se $\bar{C}'(1) = E[X(X-2)]/2$ e, por sua vez, $P'(1) = r[\bar{C}(1) + \bar{C}'(1)]/1 - r\bar{C}(1)$, para chegar-se à quantidade média no sistema,

$$L = \frac{r\{E[X] + E[X^2]\}}{2(1-\rho)} = \frac{\rho + rE[X^2]}{2(1-\rho)}.$$

$\rho < 1$ é condição necessária e suficiente para a estacionariedade.

As outras medidas de eficiência do sistema são encontradas facilmente utilizando o resultado $L_q = L - (1 - P_0) = L - \rho$ e depois a fórmulas de Little.

Além disso, podem-se estender esses resultados para o modelo $M^{[X]}/M/c$, da mesma maneira que tem-se $M/M/1$ e $M/M/c$.

8.4 Demonstrações - $M/G/1$

O processo estocástico embutido $X(t_i)$, em que X denota a quantidade no sistema e $t_1, t_2, \ldots$ são os tempos sucessivos de conclusão do serviço do i-ésimo cliente, representa o número de clientes deixados atrás do i-ésimo cliente quando este sai do sistema. De maneira mais simples, X_i é o número de clientes restantes no sistema após a saída do i-ésimo cliente.

$$X_{n+1} = \begin{cases} X_n - 1 + A_{n+1} & (X_n \geq 1) \\ A_{n+1} & (X_n = 0), \end{cases}$$

em que X_n é a quantidade no sistema no n-ésimo ponto de saída e A_{n+1} é o número de clientes que chegaram durante o tempo de serviço $S^{(n+1)}$ do cliente $(n+1)$.

Pode-se reescrever essa equação como

$$X_{n+1} = X_n - U(X_n) + A,$$

em que U é a função indicadora unitária

$$U(X_n) = \begin{cases} 1 & (X_n \geq 1) \\ 0 & (X_n = 0), \end{cases}$$

Ao calcular os valores esperados dessa variáveis aleatórias (Gross, seção 5.1.1) e notando que $E[X_{n+1}] = E[X_n] = L^{(D)}$, em que $L^{(D)}$ é o tamanho esperado do sistema em estacionariedade nos pontos de saída (*departure points*), pode-se obter o resultado

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

pois o tamanho esperado do sistema em estacionariedade nos pontos de saída é igual em qualquer ponto (obs.:$\sigma_S^2 = Var(S)$).

Pelo vetor de probabilidade do estado estacionário,

$$\pi \mathbf{P} = \pi,$$

em que,

$$\mathbf{P} = \begin{pmatrix} k_0 & k_1 & k_2 & \cdots \\ k_0 & k_1 & k_2 & \cdots \\ 0 & k_0 & k_1 & \cdots \\ 0 & 0 & k_0 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

e

$$k_n = P(n \text{ chegadas durante um tempo de serviço}) = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} dB(t)$$

em que, por sua vez, $B(t)$ é a função de distribuição acumulada dos tempos de serviço, pode-se encontrar (Gross, seção 5.1.2)

8.5 Demonstrações - $G/M/1$

De maneira similar ao modelo anterior, tem-se as equações estacionárias usuais

$$\mathbf{qP} = \mathbf{q}$$

em que

$$\mathbf{P} = \begin{pmatrix} 1-b_0 & b_0 & 0 & 0 & 0 & \cdots \\ 1-\sum_{k=0}^{1} b_k & b_1 & b_0 & 0 & 0 & \cdots \\ 1-\sum_{k=0}^{2} b_k & b_2 & b_1 & b_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

e

$$b_n = P(n \text{ atendimentos durante um tempo entre chegadas}) = \int_0^\infty \frac{e^{-\mu t}(\mu t)^n}{n!} dA(t),$$

sendo $A(t)$ a f.d.a. dos tempos entre chegada T.

O vetor $\mathbf{qP} = \mathbf{q}$ leva a

$$q_i = \sum_{k=0}^{\infty} q_{i-j+1} b_k \qquad (i \geq 1)$$
$$q_0 = \sum_{j=0}^{\infty} q_j \left(1 - \sum_{k=0}^{j} b_k\right)$$

que, após demonstrações não-triviais (Gross, seção 5.3.1), chega-se a

$$q_n = (1 - r_0) r_0^n \qquad (n \geq 0,\ \rho < 1)$$

em que r_0 é a única raiz que soluciona

$$z = A^*[\mu(1-z)],$$

sendo $A(z)$ a transformada Laplace-Stieltjes (LST) da f.d.a. dos tempos entre chegada.

A complexidade está justamente na raiz r_0, pois os próximos resultados são análogos ao modelo básico $M/M/1$, com r_0 substituindo ρ, e com a diferença de que os tamanhos médios são válidos apenas para os momentos de chegada, indicado por $L^{(A)}$.

$$L^{(A)} = \frac{r_0}{1-r_0} \qquad e \qquad L_q^{(A)} = \frac{r_0^2}{1-r_0}.$$

Exemplo teórico - G/M/1

Considerando um hospital em que as funções de distribuição acumulada dos tempos entre chegada e dos tempos de serviço são dadas, respectivamente, por:

$$A(s) = 2(e^{-s} - e^{-2s})$$
$$B(s) = e^{-s}$$

cujas transformadas ficam

$$A^*(s) = \frac{2}{(s+1)(s+2)}$$
$$B^*(s) = \frac{1}{s+1}$$

Pelo método de fatorização espectral, temos

$$\psi_+(s) = \frac{s[s-(1-\sqrt{2})]}{s+1}$$
$$\psi_-(s) = -\frac{(1-s)(2-s)}{s-(1+\sqrt{2})}$$

que utilizamos para encontrar a transformada da f.d.a. do tempo de espera

$$\bar{W}(s) = \frac{1}{s} - \frac{2-\sqrt{2}}{s-(1-\sqrt{2})}$$

que, invertendo, obtemos

$$W(t) = 1 - (2-\sqrt{2})e^{-\left(\sqrt{2}-1\right)t}, \quad t \geq 0$$

A partir disso, obtemos as medidas de performance, antes encontrando $\mu = 1$ ao notar que, na verdade, temos um problema $G/M/1$,

$$W = \sqrt{2} \qquad W_q = 0.4142136$$

Também podemos calcular a probabilidade de se esperar na fila mais que dois minutos ($t = 2$),

$$P(T_q > t) = 1 - W_q(t) = (2 - \sqrt{2})e^{-(\sqrt{2}-1)t} = 0.256.$$

Note-se que, sem o conhecimento da taxa média das chegadas ao sistema, não podemos calcular outras medidas de desempenho. Além disso, se houver maiores níveis de complexidade, como desistência, prioridade de atendimento, retorno de paciente e múltiplos processos de chegada, os modelos analíticos não servem para analisar tal sistema. É o que veremos no exemplo da subseção sobre simulação.