Processos Estocásticos - Parte V
Introdução à Teoria de Filas

Héliton R. Tavares

Universidade Federal do Pará
www.ufpa.br/heliton
heliton@ufpa.br

Índice

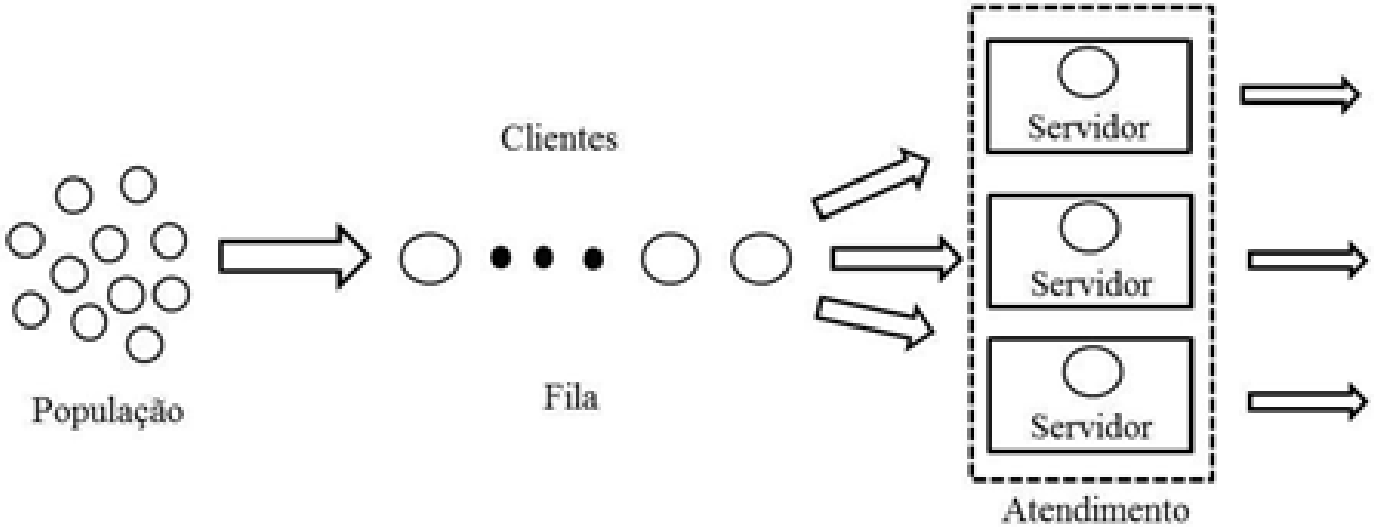
Sites e materiais de apoio

- Wikipédia PT:
  `https://pt.wikipedia.org/wiki/Teoria_das_filas`
- Wikipédia EN:
  `https://en.wikipedia.org/wiki/Queueing_theory`
- Wikipédia M/M/1:
  `https://pt.wikipedia.org/wiki/Fila_M/M/1`

Introdução

A Teoria de Filas é um ramo da probabilidade que consiste em estudar o comportamento das filas de forma analítica, visando, em geral, melhorias no desempenho do sistema, como, por exemplo, a redução do tempo em fila ou um atendimento mais rápido e eficaz. Assim, é possível tomar decisões adequadas para uma provável modificação do sistema em estudo. De modo geral, esses sistemas são formados por clientes que chegam para um determinado tipo de atendimento e, quando a capacidade de serviço é menor do que o número de clientes na chegada, estes clientes precisam esperar, ocorrendo à formação de filas.

Elementos e características de uma fila



Elementos de uma Fila

i) Modelo de Chegada: É a distribuição ou modelo determinístico das chegadas de clientes ao sistema.

ii) Modelo de Atendimento: É especificado pelo modelo, que pode ser determinístico ou estocástico, do tempo que o atendente leva para concluir o serviço.

iii) Número de Servidores: Corresponde ao número de elementos, ou atendentes, que realizam o serviço no sistema.

iv) Capacidade do Sistema: É o número máximo de clientes que o sistema é capaz de manter em espera, incluindo fila e atendimento.

v) Tamanho da População: Número de clientes que podem chegar ao sistema.

vi) Disciplina da Fila: É a forma de atendimento aos clientes. As principais são:

- FIFO (do inglês first in, first out): O primeiro a chegar é o primeiro a ser atendido.
- LIFO (do inglês last in, first out): O último a chegar é o primeiro a ser atendido.
- Aleatório/Randômico: O atendimento é realizado sem preocupação com a ordem de chegada.
- Com Prioridade: O atendimento é realizado considerando critérios estabelecidos para diferentes tipos de clientes.

Notação de Kendall

Uma fila pode ser descrita pela chamada notação de Kendall, que possui o seguinte formato: **A/B/s/K/N/Z**:

- **A** representa a distribuição dos intervalos entre chegadas;
- **B** representa a distribuição do tempo de serviço;
- **s** se refere ao número de guichês de atendimento disponíveis;
- **K** é a capacidade máxima no sistema;
- **N** é o tamanho da população
- **Z** é a disciplina da fila

Alguns autores simplificam a notação de Kendall omitindo os três últimos elementos (K/N/Z) indicando que a fila possui capacidade ilimitada de receber clientes no sistema ($K = \infty$), com uma população infinita ($N = \infty$) e disciplina Z=FIFO.

$$\mathbf{A/B/s} = \mathbf{A/B/s/\infty/\infty/FIFO}$$

Distribuições

Alguns exemplos de distribuições adotadas pelos parâmetros A e B, com suas respectivas abreviações, são:

- **M** : Memoryless/Markoviano (Exponencial (tempo) $\leftrightarrow$ Poisson (taxa));
- **D** : Determinístico (tempo constante);
- **U** : Uniforme;
- E_k : Distribuição de Erlang do tipo k ou Gama $\leftrightarrow$ soma de distrib. exponenciais independentes;
- H_k : Hiperexponencial
- **G** : distribuição geral (não se sabe nada sobre os tempos de chegada/serviço);
- **GI** : distribuição geral em que os tempos de chegada/serviço são $i.i.d.$

Casos mais comuns: M/M/1, M/M/s, M/M/1/K, M/M/s/K.

Tipos de fila

- Canal único, fase única: Um exemplo típico é um salão de beleza com uma única pessoa atendendo.
- Canal único, fases múltiplas: O sistema de lavagem de carros é uma ilustração porque uma série de serviços é realizada em uma sequência bastante uniforme. Um fator crítico é a quantidade de itens permitidos à frente de cada serviço, o que, por sua vez, constitui filas de espera separadas.
- Canais múltiplos, fase única: Os guichês nas lojas de departamentos exemplificam esse tipo de estrutura. Os diferentes tempos de serviço dedicados a cada cliente resultam em velocidade e fluxo desigual entre as filas.
- Canais múltiplos, fases múltiplas: A admissão de pacientes em um hospital segue este padrão, porque uma sequência específica de etapas é, geralmente, completada: contato inicial no balcão de admissões, preenchimento de formulários, confecção das pulseiras de identificação, obtenção de um quarto, acompanhamento do paciente até o quarto, e assim por diante.
- Misto: Consideram-se duas subcategorias:
  - Estruturas múltiplas para canais únicos: filas que se unem em uma única fila para o serviço de fase única e as filas que se juntam em uma para o serviço de fases múltiplas.
  - Estruturas de caminhos alternativos: encontram-se duas estruturas que diferem nas exigências de fluxo direcional. A primeira é similar a estrutura de canais múltiplos e fases múltiplas, com a diferença de que (a) pode haver mudança de um canal para o próximo depois que o primeiro serviço foi realizado e (b) o número de canais e fases pode variar – novamente – depois da realização do primeiro serviço.

Medidas de Eficiência do Sistema

A análise com a teoria de filas é obtida para um sistema que está em regime estacionário, ou seja, o sistema está em equilíbrio. As principais variáveis de interesse nessa circunstância são:

L : Número médio de clientes no sistema (inclui os que estão sendo atendidos);

L_q : Número médio de clientes na fila $= \sum_{n=0}^{\infty} nP_n$;

W : Tempo médio de permanência de um cliente no sistema (inclui os que estão sendo atendidos) $= \sum_{n=s}^{\infty}(n-s)P_n$;

W_q : Tempo médio de espera de um cliente na fila;

P_n : Probabilidade de que o número de clientes no sistema seja n;

μ : Taxa de atendimento;

λ : Taxa de chegadas;

$\rho = \frac{\lambda}{s\mu}$ Taxa de ocupação/utilização do sistema.

Obs: A condição básica para que um sistema de fila seja estável é que o fator de utilização seja menor que 1.

Modelo M/M/1

Considerado um modelo de fila mais simples, representado na Figura abaixo, o modelo M/M/1 possui seus processos de chegada e formas de atendimento dados por distribuição Exponencial, no qual se tem apenas um atendente.

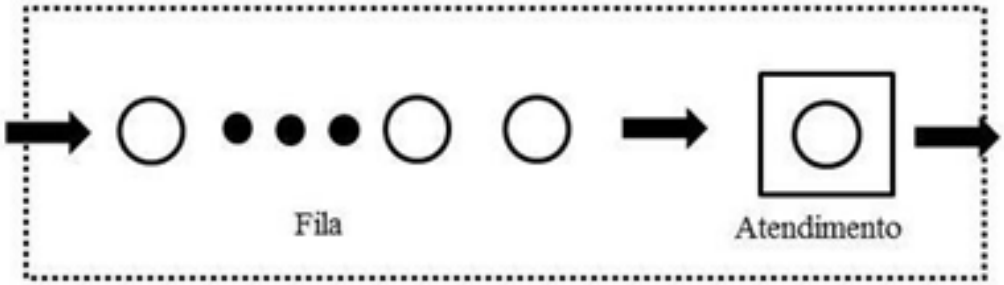


Tabela 1: Medidas de desempenho para o M/M/1

Descrição	Expressão
Número de clientes esperando no sistema	$L = \frac{\lambda}{\mu-\lambda}$
Comprimento esperado da fila	$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$
Tempo de espera no sistema	$W = \frac{1}{\mu-\lambda}$
Tempo de espera na fila	$W_q = \frac{\lambda}{\mu(\mu-\lambda)}$
Existir n clientes no sistema	$P_n = \rho^n(1-\rho)$

Barbeiro

Exemplo

Clientes chegam a uma barbearia, de um único barbeiro, com uma duração média entre chegadas de 20 minutos. O barbeiro gasta em média 15 minutos com cada cliente.
a) Qual a probabilidade de um cliente não ter que esperar para ser atendido?
b) Qual o número esperado de clientes no salão do barbeiro? E na fila?
c) Quanto tempo, em média, um cliente permanece no salão?
d) Quanto tempo, em média, um cliente espera na fila?
e) O barbeiro está estudando a possibilidade de colocar outro barbeiro desde que o tempo de permanência médio de cada cliente no salão passe a 1,25 hora. Para quanto deve aumentar a taxa de chegada de modo que este segundo barbeiro fique justificado?

Solução:

Em uma hora: Duração média entre chegadas é 20 minutos. Chegam 3 clientes por hora. O tempo de serviço do barbeiro é 15 minutos com cada cliente, logo ele atende em média 4 clientes por hora. Então:

Taxa de chegada: $\lambda = 3$ clientes/h.

Taxa de serviço: $\mu = 4$ clientes/h.

(a) Para calcularmos a probabilidade de um cliente não ter que esperar para ser servido é necessário que o sistema esteja ocioso, assim

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{3}{4} = \frac{4-3}{4} = \frac{1}{4} = 0,25 \text{ ou } 25\%$$

(b) O número esperado de clientes no salão do barbeiro é dado por:

$$L = \frac{\lambda}{\mu - \lambda} = \frac{3}{4-3} = 3 \text{ clientes.}$$

O número esperado de clientes na fila é dado por:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{3^2}{4(4-3)} = \frac{9}{4} = 2,25 \text{ clientes}$$

(c) O tempo, em média, que um cliente permanece no salão é calculado por:

$$W = \frac{1}{(\mu - \lambda)} = \frac{1}{4-3} = 1 \text{ hora}$$

(d) O tempo, em média, que um cliente espera na fila é calculado por:

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{3}{4(4-3)} = \frac{3}{4} = 0,75 \text{ hora}$$

(e) O objetivo do estudo do barbeiro leva em consideração a contratação de outro barbeiro em sua barbearia, mas para isso, tomaremos o tempo de permanência médio de cada cliente no salão de $1,25$ hora, ou seja, $W = 1,25$ h, consideraremos ainda que a taxa de serviço μ continua a mesma e devemos calcular o valor de λ. Então, como:

$$W = \frac{1}{(\mu - \lambda)}, \text{ temos } 1,25 = \frac{1}{(\mu - \lambda)} \Rightarrow 1,25 = \frac{1}{(4 - \lambda)} \Rightarrow 5 - 1,25\lambda = 1 \Rightarrow \lambda = \frac{4}{1,25} \Rightarrow \lambda = 3,2$$

Portanto, para que o barbeiro coloque outro barbeiro no salão a taxa de chegada deve aumentar em $3,2$ clientes/h.

Modelo M/M/s

O modelo M/M/s possui seus processos de chegada e formas de atendimento dados por distribuição Exponencial, porém agora com s atendentes. Na Figura abaixo temos a representação do M/M/3:

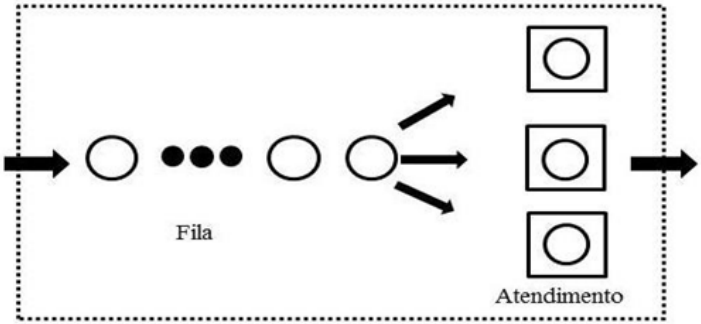


Tabela 2: Medidas de desempenho para o M/M/s

Descrição	Expressão
Número de clientes esperando no sistema	$L = L_q + \frac{\lambda}{\mu}$
Comprimento esperado da fila	$L_q = \frac{P_0(\lambda/\mu)^2\rho}{s!(1-\rho)^2}$
Tempo de espera no sistema	$W = W_q + \frac{1}{\mu}$
Tempo de espera na fila	$W_q = \frac{L_q}{\lambda}$
Existir $n = 0$ clientes no sistema	$P_0 = \sum_{n=0}^{s-1} \rho^n/n! + s\rho^s/[s!]$
Existir $n > 0$ clientes no sistema	$P_n = \begin{cases} P_0\rho^n/n! & 1 \leq n < s \\ P_0\rho^n/[s^{n-s}s!] & n \geq s \end{cases}$

Modelo M/M/1/K

O modelo M/M/1/K possui seus processos de chegada e formas de atendimento dados por distribuição Exponencial, com 1 atendente, mas agora com capacidade limitada K.

Modelo M/M/s/K

O modelo M/M/s/K possui seus processos de chegada e formas de atendimento dados por distribuição Exponencial, porém agora com s atendentes e com capacidade limitada K.

Resultados ou Lei de Little

Desenvolvida por John Little no início dos anos 60, A Lei de Little relaciona o número de clientes no sistema com o tempo médio despendido no sistema.

$$L = \lambda W \quad \text{e} \quad L_q = \lambda W_q$$

Exemplo

Numa sala de espera de um consultório, há 15 clientes em média e taxa de chegada é de 1 cliente a cada 30 segundos. Calcule o tempo médio de espera dos clientes na sala. Os clientes são atendidos na ordem de chegada (FIFO).

Temos que:
$L = 15$ e $\lambda = 2$ clientes/minuto. Aplicando a equação de Little na fila, teremos que o tempo de espera na fila será $W = 15/2 = 7,5$ minutos.