

# Uma introdução à Teoria das Filas

Aldo William Medina Garay  
Introdução aos Processos Estocásticos

13/06/2012

# Quem nunca pegou fila na vida?



**Figura:** Experiência no bandeirão

# Motivação

- As filas estão presentes em nosso cotidiano, no supermercado, no banco, no trânsito, em qualquer situação em que precisamos esperar por um serviço ou oportunidade.
- Um sistema de filas pode ser descrito como clientes que chegam para um determinado serviço em que são atendidos imediatamente ou esperam, saindo após o atendimento.
- O principal motivo de se estudar Teoria de Filas é **otimizar** o sistema, que se caracteriza por:
  - melhor utilização dos serviços disponíveis,
  - menor tempo de espera,
  - maior rapidez no atendimento.

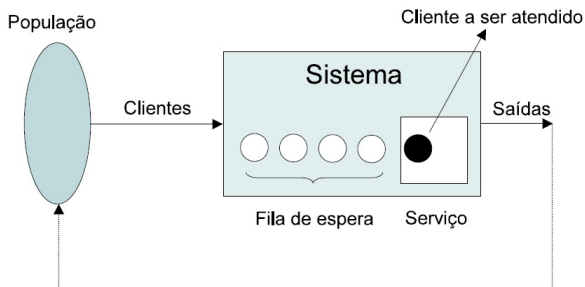
# Estrutura de um sistema de fila de espera

- **Fonte** ou **População**, que gera os clientes que vão chegar ao sistema.
- **Fila**, construída pelos clientes à espera de ser atendidos (Não inclui o(s) cliente(s) em atendimento).
- **Serviço** ou atendimento, que pode ser constituído por um ou mais postos de atendimento.

$$\text{Fila} + \text{Serviço} = \text{Sistema}$$

- Número de clientes no sistema (Em cada instante) = **Estado do sistema**.

# Estrutura de um sistema de fila de espera



**Figura:** Estrutura de um sistema de fila de espera

# Fonte

- **Dimensão da população:**
  - Infinita: quando a probabilidade de ocorrer uma nova chegada não é influenciada pelo número de clientes que já se encontram no sistema .
  - Finita
- **Dimensão da chegada:**
  - Clientes chegam um a um.
  - Clientes chegam em grupo.
- **Controle das chegadas:**
  - Chegadas controláveis (Por exemplo, inscrições em dias fixos).
  - Chegadas incontroláveis (Por exemplo, urgência de um hospital).

# Cliente

- **Distribuição das chegadas:**

- O padrão das chegadas pode ser descrito pelo tempo entre duas chegadas consecutivas (**Tempo entre chegadas**) ou pelo número de chegadas por unidade de tempo. (Distribuição das chegadas) .

- **Taxas das chegadas ( $\lambda$ ):**

- Número médio de clientes que procuram o serviço por unidade de tempo.

- **Atitude dos clientes:**

- *Paciente*, permanecem na fila até serem atendidos.

- *Impaciente*, desistem de esperar ou simplesmente não se juntam à fila se esta for muito grande.

# Fila

- **Número de filas:**

- Fila simples: uma única fila mesmo que o servidor tenha vários postos de atendimento.
- Fila múltipla: uma fila por posto de atendimento; cada posto de atendimento constitui um sistema separado de fila de espera.

- **Comprimento da fila:**

- Infinito: A capacidade máxima da fila é muito grande quando comparada com o número de elementos que habitualmente a constituem.
- Finito: A fila pode acolher apenas um número determinado (pequeno) de clientes.

- **Disciplina da fila:**

- FIFO: 'First in First Out'.
- Prioridades: Reservas, idade, emergência.



# Serviço

- **Estágios do serviço:**

- Um sistema com um único Estágio: Por exemplo uma barberia
- Um sistema de multi-estágio

- **Dimensão do serviço:**

- Simples;
- Em grupo (Por exemplo um elevador atende vários clientes simultaneamente).

- **Distribuição do tempo de serviço:**

- Constante
- Aleatório: Distribuição exponencial, Erlang, entre outras.

- **Taxa de serviço ( $\mu$ ):**

- Número médio de clientes que podem ser atendidos por cada servidor por unidade de tempo.  $\frac{1}{\mu}$  é a duração média do serviço.

# Medidas de desempenho

- Comprimento médio da fila ( $L_q$ ).
- Número médio de clientes no sistema ( $L$ ).
- Tempo médio de espera da fila ( $W_q$ ).
- Tempo médio de espera no sistema ( $W$ ).
- Tempo médio de ocupação (e desocupação) do serviço (percentagem de tempo durante o qual o serviço está ocupado).

# Notação no sistema das filas: $X/Y/Z/W$

- A notação de processos de filas mais utilizada atualmente foi proposta por Kendall, em 1953.
- $X, Y$  representam as distribuições do intervalo de tempo entre chegadas e do tempo de serviço respectivamente, onde:
  - $M$  representa a distribuição exponencial,
  - $G$  representa uma distribuição não especificada,
  - $D$  representa as chegadas ou atendimentos determinísticos.
- $Z$ , representa o número de servidores em paralelo.
- $W$ , representa outras características do sistema, tais como comprimento da fila.
- Entre os tipos de filas mais conhecidos temos o  $M/M/1$ ,  $M/M/c$ ,  $M/M/\infty$ , entre outros.

# Tipo $M/M/1$

- Suponha que os clientes chegam a uma estação de um único servidor, de acordo com um processo de Poisson com taxa de  $\lambda$ .
- Após a chegada, cada cliente se encaminha diretamente ao serviço se o servidor está livre, caso contrário ele espera na fila.
- Quando o servidor termina de atender um cliente, o cliente deixa o sistema e o próximo cliente da fila será atendido.
- Os tempos sucessivos de serviço de atendimento dos clientes são considerados variáveis aleatórias exponenciais independentes com média  $1/\mu$ .
- Seja  $X(t)$  o número de clientes no sistema no tempo  $t$ , então  $\{X(t), t \geq 0\}$ .

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

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

# Tipo $M/M/1$

- Distribuição Estacionária**

Seja  $\pi_n(t) = P(N(t) = n)$ , pode-se mostrar que:

$$\pi_{n+1} = -\frac{(\lambda + \mu)}{\mu} \pi_n + \pi_{n+1} + \frac{\lambda}{\mu} \pi_{n-1}, \quad n \geq 1$$
$$\pi_1 = \frac{\lambda}{\mu} \pi_0$$

Se  $\lambda > \mu$  o número médio de chegadas por unidade de tempo é maior que o número médio de saídas por unidade de tempo, e  $X_t \rightarrow \infty$  quase certamente.

Se  $\lambda = \mu$  o processo é recorrente nulo.

Se  $\lambda < \mu$ , o processo é recorrente positivo, e considerando  $\sum_{n=1}^{\infty} \pi_n = 1$ , pode-se mostrar por indução que a distribuição estacionária é dada por:

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n \quad (1)$$

# Tipo $M/M/1$

- **Valor esperado do número de clientes**

Sob condição de equilíbrio, seja  $N$  o número total de usuários no sistema, temos que seu valor esperado é:

$$L = E\pi(X_t) = E(N) = \sum_{n=0}^{\infty} n\pi_n = \sum_{n=0}^{\infty} n \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n = \frac{\lambda}{\mu - \lambda} \quad (2)$$

Sob condição de equilíbrio, seja  $N_q$  o número total de usuários na fila, temos que seu valor esperado é:

$$L_q = E(N_q) = \sum_{n=1}^{\infty} (n-1)\pi_n = \sum_{n=1}^{\infty} n\pi_n - \sum_{n=1}^{\infty} \pi_n = \frac{\lambda}{\mu} \left(\frac{\lambda}{\mu - \lambda}\right) \quad (3)$$

# Tipo $M/M/1$

- **Medida de desempenho do sistema**

Relacionar o número médio de usuários na fila ou no sistema com o tempo médio de espera na fila denotadas por  $W$  e  $W_q$  respectivamente.

Essas relações, que foram apresentadas por Little em 1961, são dadas pelas seguintes expressões:

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

# Tipo $M/M/1$

- Exemplo**

Seja um aeroporto com uma única pista de pouso/decolagem. Os aviões chegam a uma taxa de 15/hora, e levam em média 3 minutos para aterrisar. Assumindo que as chegadas são um processo de Poisson, e o tempo de aterrisagem é distribuído por uma exponencial.

$$\lambda = 15/hora \quad \text{e} \quad \mu = \frac{60}{3}/hora = 20/hora$$

$$\text{Intensidade de tráfego: } = \frac{\lambda}{\mu} = \frac{3}{4} = 0.75$$

$$\text{Número médio de aviões aguardando para pousar: } E(N_q) = \frac{\lambda}{\mu} * \frac{\lambda}{(\mu - \lambda)} = 2.25$$

$$\text{Tempo médio de espera para o pouso: } W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{3}{20} = 9 \text{ minutos}$$



# Tipo $M/M/c$

- É uma extensão da fila  $M/M/1$ , apresentando a mesma distribuição de chegada, a mesma disciplina e a mesma distribuição do tempo de serviço. O que difere de ambas é a quantidade de servidores, agora considerado um sistema com  $c$  servidores.
- Agora as taxas de chegadas e saídas são dadas por:

$$\lambda_n = \lambda, \quad n \geq 0$$
$$\mu_n = \{n\mu, \quad 0 \leq n < c \quad \text{e} \quad c\mu, \quad n \geq c\}$$

# Tipo $M/M/c$

- Distribuição Estacionária**

Consideramos agora que a intensidade de tráfego é dada por:  $\frac{\lambda}{c\mu}$ ,  $n \geq c$

Com o fato do estado se encontrar em equilíbrio, pode-se mostrar que

$$\pi_n = \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n \pi_0 \quad 0 \leq n < c; \quad \text{e} \quad \frac{1}{c!c^{n-c}} \left( \frac{\lambda}{\mu} \right)^n \pi_0 \quad n \geq c \quad (4)$$

Nesta fila também temos que  $\lambda < c\mu$ , neste caso o processo também será recorrente positivo.

# Tipo $M/M/c$

- **Valor esperado do número de clientes**

Sob condição de equilíbrio, o valor esperado do número de pessoas na fila é dada por:

$$E(N_q) = \sum_{n=c}^{\infty} (n-c) \pi_n = \frac{1}{(c-1)!} \left(\frac{\lambda}{\mu}\right)^c \pi_0 \frac{\lambda\mu}{(c\mu - \lambda)^2} \quad (5)$$

- **Medida de desempenho do sistema**

Utilizando a relação de Little, obtemos que

$$W_q = \frac{E(N_q)}{\lambda} = \frac{1}{(c-1)!} \left(\frac{\lambda}{\mu}\right)^c \pi_0 \frac{\mu}{(c\mu - \lambda)^2}$$

# Tipo $M/M/c$

- Exemplo**

Reconsiderando o exemplo utilizado para o tipo de filas  $M/M/1$ , seja um aeroporto, agora com duas pistas de pouso/decolagem. Os aviões chegam a uma taxa de 15/hora, e levam em média 3 minutos para aterrissar. Assumindo que as chegadas são um processo de Poisson, e o tempo de aterrissagem seguem uma distribuição exponencial.

$$\lambda = 15/\text{hora}, \quad \mu = \frac{60}{3}/\text{hora} = 20/\text{hora} \quad \text{e} \quad \text{o número de servidores: } c = 2$$

$$\text{Intensidade de tráfego: } = \frac{\lambda}{c\mu} = \frac{3}{8} = 0.375$$

# Tipo $M/M/c$

- Exemplo (Continuação)**

Número médio de aviões aguardando para pousar:

$$E(N_q) = \frac{(\lambda/c\mu) (\lambda/\mu)^c \pi_0}{c! (1 - \lambda/c\mu)^2} = \frac{(3/8) * (3/4)^2 0.4545}{2 (5/8)^2} = 0.1227$$

Tempo médio de espera para o pouso:

$$W_q = \frac{(\lambda/\mu)^c \pi_0}{c! c\mu (1 - \lambda/c\mu)^2} = \frac{3}{20} = \frac{(3/4)^2 0.4545}{2 * 2 * 20 (5/8)^2} = 0.49, \text{ minutos}$$

# CASO DE ESTUDO: CINEMA

Um grupo de empresários estão reorganizando o atendimento ao público em um cinema que eles abriram há 6 meses.

Nesta primeira etapa da reorganização eles querem saber quantos caixas deveriam contratar para que maximizem seus ganhos, num total máximo de 6 atendentes, considerando:

- O cinema funciona diariamente no horário das 20:00-24:00 horas. O cinema tem 5 salas, com uma capacidade total de 720 pessoas, mas os proprietários sabem que todos os dias a quantidade máxima de pessoas que chega no local é 700.

# CASO DE ESTUDO: CINEMA

- O número de chegadas segue aproximadamente uma distribuição Poisson, de parâmetro  $\lambda = 2.50$  pessoas por minuto, e os tempos de atendimento aos clientes seguem uma distribuição exponencial com taxa  $\mu = 1.5$  atendimentos por minuto.
- O custo por hora de cada atendente é de R\$25.00.  
Para o preço do ingresso foi considerado o seguinte esquema: Ao entrar na fila de atendimento, cada cliente recebe uma senha com a hora da entrada. Ao ser atendido, o preço do ingresso, por pessoa, será de R\$20.00 – R\$0.25 \*  $T$ , onde  $T$  representa o tempo de espera na fila, em minutos.

# References

- Gross, D., Harris. C. M. Fundamentals of Queueing Theory. Willey series in Probability and Mathematical Statistics, New York 1974.
- Pardoux Etienne, Markov Processes and Applications, 2008. Wiley- Dunod Series.
- Ross, M. Sheldon, Introduction to Probability Models, 2003. Academic Press.