2. Filas de espera – introdução

Uma situação de espera ocorre quando a procura de recursos/serviços excede a disponibilidade do sistema.

As situações de espera são muito comuns no nosso dia-a-dia. Eis alguns exemplos típicos:

- Bancos/supermercados – espera pelo serviço de caixa,
- Computadores – espera por uma resposta,
- Trânsito – espera em semáforos e cruzamentos,
- Transporte público – espera nas estações e apeadeiros.

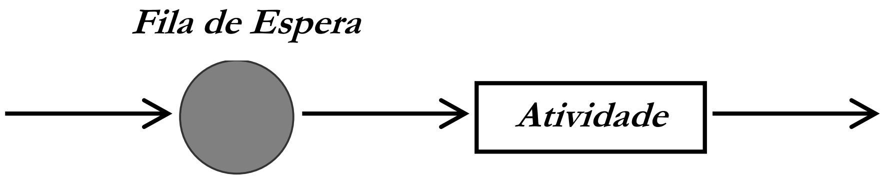
O problema consiste em encontrar uma solução que utilize um adequado nível de recursos para manter o sistema em operação económica. *»» Balanço entre o nível de serviço prestado aos clientes (“pequenas filas requerem muitos servidores”) e considerações económicas (“não demasiados servidores”).*

Exemplo: Chegada de barcos a um porto

- Cliente: Barco
- Serviço: Carga e descarga
- Facilidades: Cais de acostagem, guindastes
- Fator de variabilidade: Condições atmosféricas
- Fatores económicos: A paragem dos barcos (à espera de atendimento) é dispendiosa; o investimento em novas facilidades também pode ser volumoso.

Elementos básicos de um sistema de espera

Essencialmente, todos os sistemas de filas de espera podem ser reduzidos a subsistemas individuais consistindo em entidades (ou clientes) à espera por determinada atividade (ou serviço):



Três elementos básicos:

(1) Processo de chegada – o modo pelo qual os clientes chegam ao sistema;

(2) Disciplina da (fila) de espera – o modo como os clientes esperam até serem atendidos/servidos, (em Quantas filas? que Tipo de prioridade?)

(3) Mecanismo de serviço.

Processo de chegada (tipos de variações)

- Número de clientes potenciais
  - Finito ou infinito.

- Número de clientes que chegam simultaneamente
  - Único,
  - Em grupos de tamanho constante ou variável.

- Intervalo de tempo entre duas chegadas consecutivas
  - Constante,
  - Completamente aleatório (Poisson),
  - Outro tipo de distribuição estatística.

- Taxa média de chegadas
  - Constante,
  - Variável com o tempo,
  - Influenciada pelo estado da fila.

- Influência externa
  - Nenhuma,
  - Resultante do desempenho de um sistema a montante.

Disciplina da(s) fila(s) *(Quantas?):*

- **FIFO** (*First-In*, *First-Out*) – Os clientes são servidos pela ordem de chegada. O primeiro a chegar é o primeiro a ser atendido.
  »» Exemplo: Caixa de Supermercado.

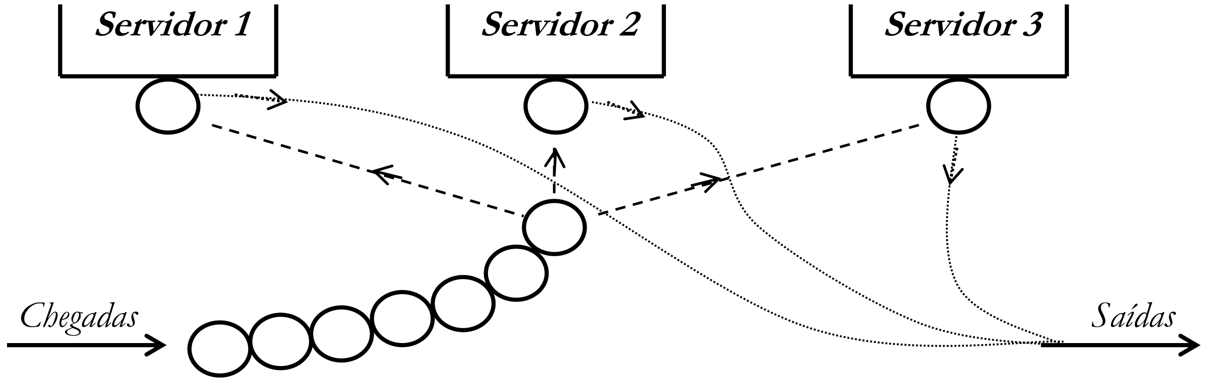
- **LIFO** (*Last-In*, *First-Out*) – Os clientes são servidos pela ordem inversa à de chegada. O último a chegar é o primeiro a ser atendido. (Conceito de "*stack*", pilha).
  »» Exemplo: Carga no porão de um navio.

- **Prioridade** – Cada cliente tem associada uma prioridade específica.
  »» Exemplo: prioridade atribuída aos pacientes num serviço de urgência de um hospital.

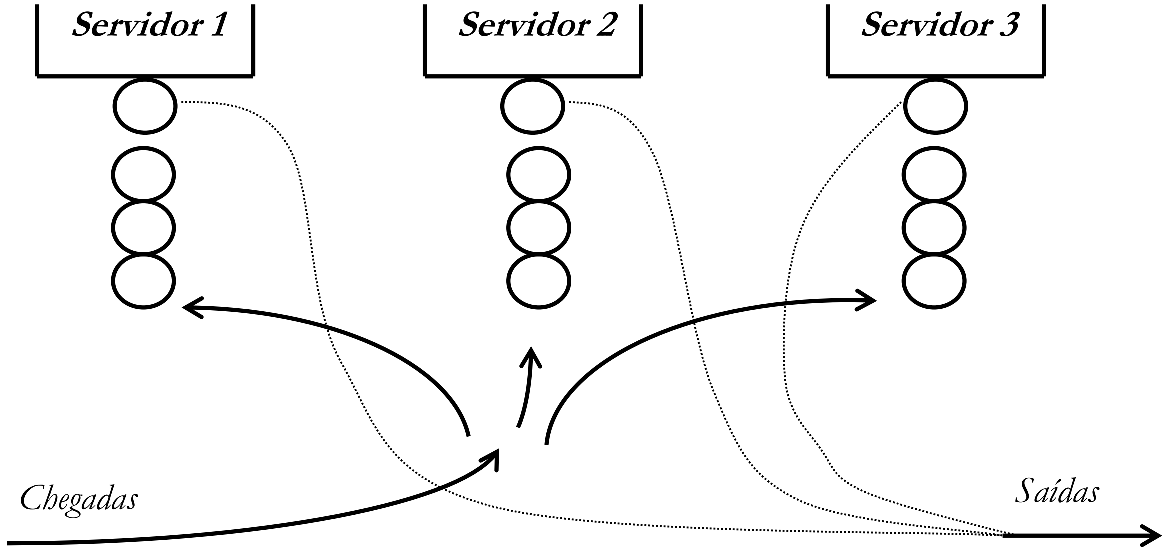
- **Remoção** ("*pre-emptive*") – O cliente que chega é colocado à frente, removendo qq cliente que lá se encontre.
  »» Exemplo: chamada de emergência no serviço de urgência de um hospital.

- **Prioridade aleatória**.

Quantas filas?



Fila de espera única



Três filas de espera independentes

Mecanismo de serviço (tipos de variações)

- Número de servidores
    - Um, vários ou variável.

- Número de clientes servidos simultaneamente
    - Um de cada vez,
    - Em grupos de tamanho constante ou variável.

- Disponibilidade do serviço
    - Permanente ou intermitente.

- Duração do serviço
    - Constante,
    - Distribuído exponencialmente (instantes de começo e finalização distribuídos independente e aleatoriamente),
    - Outras distribuições estatísticas,
    - Dependente do tempo despendido na fila.

- Taxa média de serviço
    - Constante,
    - Variável no tempo ou com o estado do sistema.

Medidas de desempenho mais comuns

- Medidas relativas a médias:
  - Número de clientes no sistema,
  - Número de clientes na fila (comprimento da fila),
  - Tempo de espera.
- Medidas relativas a probabilidades:
  - De haver *n* clientes no sistema,
  - De haver mais do que *n* clientes no sistema,
  - De se ter um tempo de espera entre *t* e *t+dt*,
  - De se ter um tempo de espera maior do que *t*,
  - De os canais de serviço estarem inativos,
  - De se ter um período de ocupação de um serviço superior a *t* unidades de tempo.

Notação de *Kendall*

$A/B/S(d\,|\,e)$

- *A* – distribuição de probabilidade relativa ao intervalo de tempo entre chegadas ao sistema,

- *B* – distribuição de probabilidade relativa às durações de serviço,

- *S* – Número de canais de serviço,

- *d* – Número máximo de clientes permitidos no sistema em qualquer instante,

- *e* – Disciplina associada à fila de espera.

Exemplos:

»» $M/M/1(\infty\,|\,FIFO)$ ou $M/M/1$

(*M* significa exponencial negativa)

»» $D/M/2$, $G/M/2(100\,|\,LIFO)$

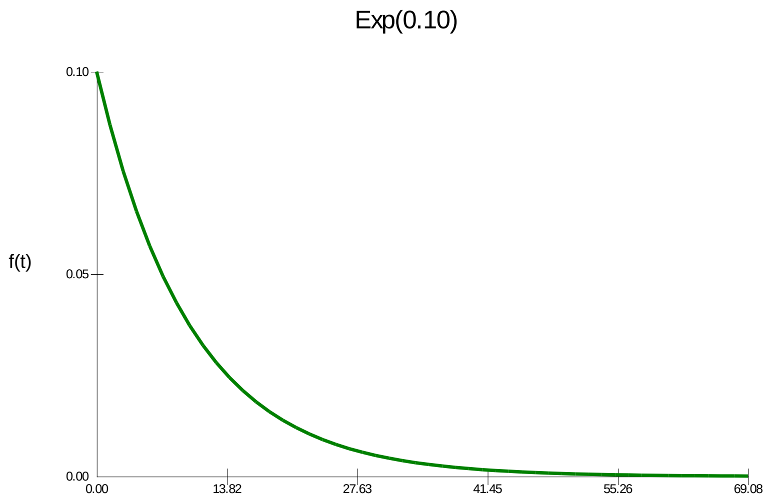
(*D* significa constante, *G* significa distribuição genérica)

Características da distribuição Exponencial Negativa

Função densidade de probabilidade (FDP):

$$f(t) = \begin{cases} \alpha e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}$$

α é o chamado parâmetro da distribuição, ex. $\alpha = 0.10$:



Função cumulativa de probabilidade:

$$F(t) = P(T \leq t) = 1 - e^{-\alpha t} \qquad (t \geq 0)$$

Média: $$E(T) = \frac{1}{\alpha}$$

Variância: $$Var(T) = \frac{1}{\alpha^2}$$

Características da distribuição de Poisson

Função densidade de probabilidade (FDP):

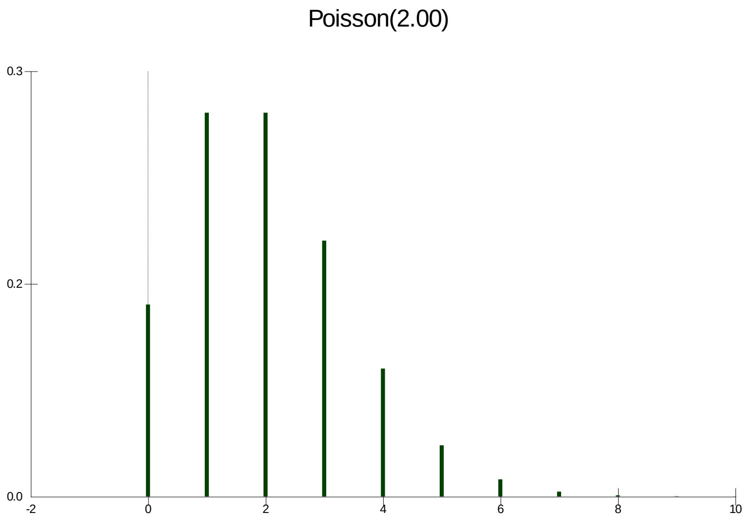
$$f(k) = \frac{(\alpha t)^k e^{-\alpha t}}{k!} \qquad (k = 0,1,2,\ldots)$$

α nº médio de acontecimentos por unidade de tempo (taxa)

t intervalo de tempo

k nº de acontecimentos durante o intervalo t

αt é o chamado parâmetro da distribuição, por exemplo $\alpha t = 2$:



Média: $E(X) = \alpha t$

Variância: $Var(X) = \alpha t$

Fila de espera $M/M/1$ (*"fila de espera simples"*)

»» Definições

Para as chegadas:

- Intervalo médio entre chegadas = *a*
- Taxa de chegada de clientes = λ (nº de clientes / u. tempo)

• $\lambda = 1/a$

- Função Densidade de Probabilidade (FDP) é $f(t) = \lambda e^{-\lambda t}$

Analogamente, para os serviços:

- Duração média de serviço = *s*
- Taxa de serviço = μ

• $\mu = 1/s$

- FDP é $g(t) = \mu e^{-\mu t}$

Fila de espera $M/M/1$

»» Taxa de chegadas

Seja $P_n(t)$ a probabilidade de haver *n* clientes no sistema no instante *t*.

Se $n > 0$, então haverá $n-1$ clientes à espera de ser atendidos (na fila) e 1 cliente a ser atendido.

Sendo o intervalo entre chegadas ao sistema uma variável aleatória que segue uma distribuição exponencial negativa, então o número de chegadas no intervalo de tempo *dt* segue uma distribuição de Poisson de média $\lambda \cdot dt$:

$$P_n = \frac{(\lambda \cdot dt)^n}{n!} e^{\lambda dt}$$

$$\begin{aligned} P_1 &= (\lambda \cdot dt) \cdot e^{\lambda dt} \\ &= (\lambda dt) \cdot \left(1 - \lambda dt + \frac{(\lambda dt)^2}{2!} + \ldots \right) \end{aligned}$$

$\boxed{\approx \quad \lambda \cdot dt \qquad \text{(para } dt \text{ suficientemente pequeno)}}$

Fila de espera $M/M/1$

»» Equações de probabilidade (1)

Considerando *dt* suficientemente curto para não haver mais do que uma chegada, nem mais do que uma saída do sistema, resulta o seguinte quadro de transições possíveis *t* → *t+dt*:

Estado no instante *t*	n $P_n(t)$	n $P_n(t)$	$n+1$ $P_{n+1}(t)$	$n-1$ $P_{n-1}(t)$
Chegadas ocorridas em *dt*	0 $1-\lambda dt$	1 λdt	0 $1-\lambda dt$	1 λdt
Serviços terminados em *dt*	0 $1-\mu dt$	1 μdt	1 μdt	0 $1-\mu dt$

$$
\begin{aligned}
P_n(t+dt) = \quad & P_n(t)\cdot(1-\lambda dt)\cdot(1-\mu dt) \\
+ \ & P_n(t)\cdot \lambda dt \cdot \mu dt \\
+ \ & P_{n+1}(t)\cdot(1-\lambda dt)\cdot \mu dt \\
+ \ & P_{n-1}(t)\cdot \lambda dt \cdot (1-\mu dt)
\end{aligned}
$$

$$
\begin{aligned}
P_n(t+dt) = \quad & P_n(t)\cdot\left(1-\lambda dt-\mu dt+\lambda\cdot\mu\cdot dt^2\right) \\
+ \ & P_n(t)\cdot \lambda\cdot\mu\cdot dt^2 \\
+ \ & P_{n+1}(t)\cdot\left(\mu dt-\lambda\cdot\mu\cdot dt^2\right) \\
+ \ & P_{n-1}(t)\cdot\left(\lambda dt-\lambda\cdot\mu\cdot dt^2\right)
\end{aligned}
$$

Fila de espera $M/M/1$

»» Equações de probabilidade (2)

Desprezando os termos em dt^2, obtém-se:

$$\boxed{P_n(t+dt)=P_n(t)\cdot(1-\lambda dt-\mu dt)+P_{n+1}(t)\cdot\mu dt+P_{n-1}(t)\cdot\lambda dt}$$

Em particular, para $n=0$, obtém-se:

$$P_0(t+dt)=P_0(t)\cdot(1-\lambda dt)+P_1(t)\cdot\mu dt$$

Podemos desenvolver a equação acima no sentido de obter:

$$\frac{P_n(t+dt)-P_n(t)}{dt}=-P_n(t)\cdot(\lambda+\mu)+P_{n+1}(t)\cdot\mu+P_{n-1}(t)\cdot\lambda$$

(N.B. $\lim\limits_{dt\to 0}\frac{P_n(t+dt)-P_n(t)}{dt}=\frac{dP_n(t)}{dt}$)

Então fica:

$$\frac{dP_n(t)}{dt}=-(\lambda+\mu)\cdot P_n(t)+\mu\cdot P_{n+1}(t)+\lambda\cdot P_{n-1}(t) \qquad (n>0)$$

$$\frac{dP_0(t)}{dt}=\mu\cdot P_1(t)-\lambda\cdot P_0(t) \qquad (n=0)$$

Fila de espera $M/M/1$

»» Comportamento a longo prazo (estado estacionário) (1)

A solução geral das equações anteriores é do tipo:

$$P_n(t) = P_n + \sum_{k=1}^{\infty} A_{n,k} e^{-C_k t}$$

sendo P_n, $A_{n,k}$ e C_k constantes ($C_k \geq 0$ para $A_{n,k} \neq 0$).

Quando $t \to \infty$, o termo exponencial tornar-se-á desprezável, e então as probabilidades deverão convergir para a probabilidade invariante P_n – diz-se que o processo atingiu o estado estacionário.

[Contudo, o nº de clientes no sistema continuará a variar!]

Estado estacionário: $\dfrac{dP_n(t)}{dt} = 0$

$$-(\lambda + \mu) \cdot P_n + \mu \cdot P_{n+1} + \lambda \cdot P_{n-1} = 0 \qquad (n = 1, 2, \ldots, \infty)$$

$$\mu \cdot P_1 - \lambda \cdot P_0 = 0 \qquad (n = 0)$$

Fila de espera $M/M/1$

»» Comportamento a longo prazo (estado estacionário) (2)

$$-(\lambda+\mu)\cdot P_n+\mu\cdot P_{n+1}+\lambda\cdot P_{n-1}=0 \qquad (n=1,2,\ldots,\infty)$$

$$\mu\cdot P_1-\lambda\cdot P_0=0 \qquad (n=0)$$

Da primeira equação tira-se sucessivamente que:

- $P_1=(\lambda/\mu)\cdot P_0$
- $P_2=(\lambda/\mu)^2\cdot P_0$
- ....
- $\boxed{P_n=(\lambda/\mu)^n\cdot P_0}$

Ao somar todas as probabilidades, obtém-se:

$$\sum_{n=0}^{\infty}P_n=P_0\cdot\sum_{n=0}^{\infty}(\lambda/\mu)^n=P_0\cdot\frac{1}{1-\lambda/\mu}$$

(com o pressuposto de que $(\lambda/\mu)=\rho<1$)

Mas como $\sum_{n=0}^{\infty}P_n=1$, vem que $P_0\cdot\frac{1}{1-\lambda/\mu}=1\boxed{\Leftrightarrow P_0=1-\rho}$

Finalmente, substituindo na equação acima enquadrada, fica:

$$\boxed{P_n=\rho^n\cdot(1-\rho)}$$

Fila de espera $M/M/1$

»» Medidas de desempenho (1)

- Utilização do servidor ou taxa de ocupação:

$$1 - P_0 = 1 - (1-\rho) = \rho$$

(Probabilidade de o servidor estar ocupado)

  - $\rho \geq 70\%$ são típicos de sistemas congestionados.
  - $\rho \geq 90\%$ não são geralmente aceitáveis na prática.

- Número médio de clientes no sistema (L):

$$L = \sum_{n=0}^{\infty} n \cdot P_n = 1 \times P_1 + 2 \times P_2 + \ldots$$

$$= \rho \cdot (1-\rho) + 2\rho^2 (1-\rho) + \ldots$$

$$= \rho \cdot (1-\rho)(1 + 2\rho + \ldots)$$

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

$$\boxed{= \frac{\rho}{(1-\rho)}}$$

Fila de espera $M/M/1$

»» Medidas de desempenho (2)

- Número médio de clientes na fila (L_q):

  (considerando que existe alguém no sistema)

$$L_q = \sum_{n=1}^{\infty}(n-1)\cdot P_n = 1\times P_2 + 2\times P_3 + \dots$$

$$= \rho^2(1-\rho) + 2\rho^3(1-\rho) + \dots$$

$$= \rho^2\cdot(1-\rho)(1+2\rho+\dots)$$

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

$$\boxed{= \frac{\rho^2}{(1-\rho)}}$$

- Probabilidade de haver mais do que *k* clientes no sistema:

$$P(n>k) = \sum_{n=k+1}^{\infty}(1-\rho)\cdot\rho^n$$

$$= (1-\rho)\rho^{k+1}(1+\rho+\rho^2+\dots)$$

$$\boxed{= \rho^{k+1}}$$

Fila de espera $M/M/1$

»» Medidas de desempenho (3)

- Probabilidade de haver mais do que *k* clientes na fila:

$$P(n > k+1) = \sum_{n=k+2}^{\infty} (1-\rho) \cdot \rho^n$$

$$= (1-\rho)\rho^{k+2}\left(1+\rho+\rho^2+\ldots\right)$$

$$\boxed{= \rho^{k+2}}$$

- Número médio na fila durante todo o tempo:

$$(1-\rho)\times 0 + \rho\left(\frac{\rho^2}{1-\rho}\right) = \boxed{\frac{\rho^3}{1-\rho}}$$

Fila de espera $M/M/1$

»» Exemplo 1

Uma ponte rolante tem um tempo médio de elevação de 10 minutos, seguindo uma distribuição exponencial negativa.
Este serviço é solicitado, em média, 4 vezes por hora. A procura ocorre aleatoriamente, seguindo também uma distribuição exponencial negativa.

($\lambda = 4/hora$, $\mu = 6/hora$)

a) Utilização do servidor: $\rho = \lambda/\mu = 66\%$

b) Nº médio no sistema: $\dfrac{\rho}{1-\rho} = 2$

c) Nº médio na fila (desde que esta exista): $\dfrac{\rho^2}{1-\rho} = 1.33$

d) Probabilidade de haver mais do que 5 clientes no sistema:
$\rho^{k+1} = \rho^6 = 0.088$

e) Probabilidade de haver mais do que 5 clientes na fila:
$\rho^{k+2} = \rho^7 = 0.059$

f) Nº médio na fila durante todo o tempo: $\dfrac{\rho^3}{1-\rho} = 0.888$

Relações fundamentais de *Little*

Estas relações são extremamente úteis, pois permitem determinar as quatro medidas de desempenho a partir da expressão analítica de uma delas.

L Número (médio) de elementos no sistema,

L_q Número (médio) de elementos na fila,

W Tempo (médio) de espera/permanência no sistema,

W_q Tempo (médio) de espera na fila.

Admitindo uma taxa de chegadas constante (λ), as relações são:

- $L = \lambda W$ (ex, 10 clientes/h, cada demora 0.2h --> L=2 clientes em média no sistema)
- $L_q = \lambda W_q$
- $W = W_q + \dfrac{1}{\mu}$ (válida qq que seja o nº de servidores, desde que a taxa de serviço seja igual para todos)
- $L = L_q + \dfrac{\lambda}{\mu}$ (deduz-se das anteriores...)

Filas de espera com comprimento limitado (*K*)

(ver págs. 76-77)

Devido a fatores diversos, tais como:

- Limitações físicas (*ex.*, bomba de gasolina na via pública);
- Limitações administrativas (*ex.*, estação de serviço/oficina).

Tipicamente, a taxa de chegadas depende do estado do sistema:

$$\lambda_n = \begin{cases} \lambda, & \text{se } n = 0,1,...,K-1 \\ 0, & \text{se } n \geq K \end{cases}$$

Então, a taxa média (ponderada) é: $\bar{\lambda} = \sum_{n=0}^{K-1} \lambda P_n = \lambda\left(1 - P_K\right)$

<u>Importante:</u> As relações fundamentais (de *Little*) mantêm-se válidas se se substituir λ por $\bar{\lambda}$.

Neste caso, $\rho = \dfrac{\lambda}{\mu}$ representa a "taxa de pressão", e o equilíbrio é possível mesmo que $\rho > 1$. (Porquê?)

- $\dfrac{\bar{\lambda}}{\mu}$ representa a taxa de ocupação, e
- P_K representa também a probabilidade de desistência por falta de capacidade do sistema!

<u>Sistemas de filas de espera com população finita (*N*)</u>

(ver págs. 78-79)

Por exemplo, situações industriais em que existe um pequeno número de máquinas sujeitas a avarias mais ou menos frequentes. Nestes casos, os clientes representam as máquinas, enquanto os servidores representam os elementos da equipa de manutenção/reparação.

A taxa de chegadas depende do estado do sistema:

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{se } n = 0,1,\dots,N-1 \\ 0, & \text{se } n \geq N \end{cases}$$

sendo, neste caso, λ a taxa de chegadas "unitária" (i.e., a taxa de chegada(s) quando existe apenas uma máquina não avariada).

Então, a taxa média (ponderada) de chegadas é:

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{N} (N-n)\lambda P_n = \lambda(N-L)$$

Como $\lambda_n = 0$ quando $n = N$, o sistema atingirá sempre o estado de equilíbrio qualquer que seja o valor de $\rho = \dfrac{\lambda}{\mu}$.

Processos de vida e morte (*M/M/1*)

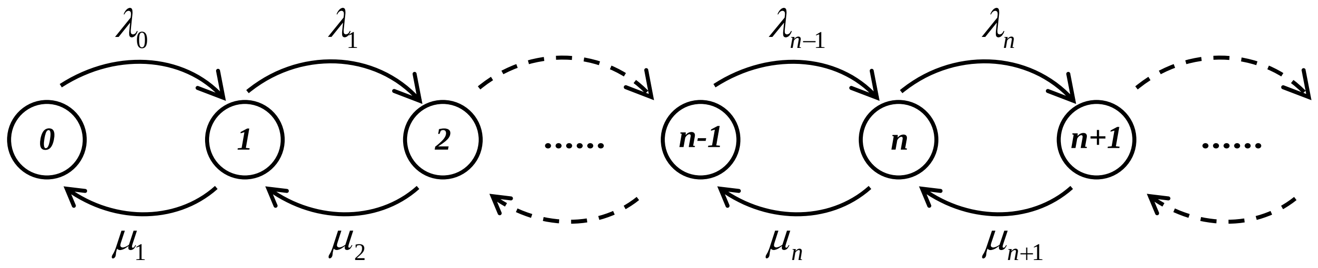


Diagrama do processo de vida e morte (genérico)

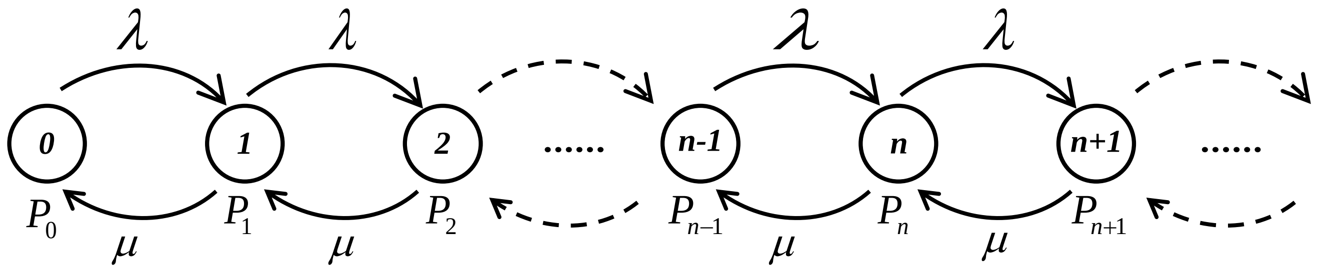


Diagrama do processo de vida e morte p/ processo (*M/M/1*)

Processos de vida e morte (*M/M/S*)

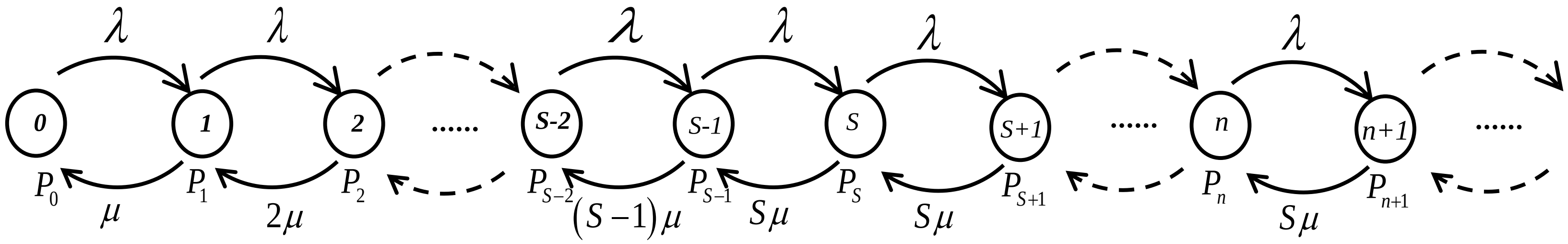


Diagrama do processo de vida e morte p/ processo (*M/M/S*)

Características do modelo ($M/M/1$) –> *"fila de espera simples"*

Chegada: Poisson

Taxa = λ clientes / u. tempo

População = ∞

Fila máxima = ∞

Tempo de atendimento: Exponencial Negativo

Taxa = μ clientes / u. tempo

Nº servidores = 1

Taxa de ocupação = $\rho = (\lambda/\mu) < 1$

Taxa de desocupação = $1-\rho$

$$L_q = \sum_{n=1}^{\infty}(n-1)P_n = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho}$$

$$L = \sum_{n=0}^{\infty} nP_n = L_q + \frac{\lambda}{\mu} = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$$

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

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

$P_0 = 1-\rho$ (taxa de desocupação)

$P_n = \rho^n P_0$

$P(n > k) = \rho^{k+1}$

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

$P(W_q > t) = \rho e^{-\mu(1-\rho)t}, \ t \geq 0$

$P(W_q = 0) = P_0$

(Fonte deste resumo e dos seguintes: L.Valadares Tavares et al., Investigação Operacional, McGraw-Hill, 1996).

Características do modelo ($M/M/S$) –> *“fila de espera simples, com S servidores”*

Chegada: Poisson

Taxa = λ clientes / u. tempo

População = ∞

Fila máxima = ∞

Tempo de atendimento: Exponencial Negativo

Taxa = μ clientes / u. tempo, servidor

Nº servidores = S

Taxa de ocupação = $\rho = \dfrac{\lambda}{S\mu} < 1$

Taxa de desocupação = $1-\rho$

$$L_q = \sum_{n=S}^{\infty}(n-S)P_n = \frac{P_0(\lambda/\mu)^S \rho}{S!(1-\rho)^2}$$

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

$$L = \sum_{n=0}^{\infty} nP_n = L_q + \frac{\lambda}{\mu}$$

$$W = W_q + \frac{1}{\mu} = \frac{L}{\lambda}$$

$$P_0 = 1 \Bigg/ \left[\sum_{n=0}^{S-1}\frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^S}{S!}\frac{1}{1-\rho}\right]$$

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!}P_0, & \text{se } 0 \le n \le S \\[2ex] \dfrac{(\lambda/\mu)^n}{S!S^{n-S}}P_0, & \text{se } n \ge S \end{cases}$$

$$P(W > t) = e^{-\mu t}\left[1 + \frac{P_0(\lambda/\mu)^S}{S!(1-\rho)}\left(\frac{1-e^{-\mu t(S-1-\lambda/\mu)}}{S-1-\lambda/\mu}\right)\right], \; t \ge 0$$

$$P(W_q > t) = \left[1 - P(W_q = 0)\right]e^{-S\mu(1-\rho)t}, \; t \ge 0$$

$$P(W_q = 0) = \sum_{n=0}^{S-1} P_n$$

Características do modelo ($M/M/1(K)$) –> *"fila de espera c/ comprimento limitado"*

Chegada: Poisson

Taxa = λ clientes / u. tempo

$$\lambda_n = \begin{cases} \lambda, & \text{se } n = 0,1,\ldots,K-1 \\ 0, & \text{se } n \geq K \end{cases}$$

$$\bar{\lambda} = \sum_{n=0}^{K-1} \lambda P_n = \lambda\left(1 - P_K\right)$$

População = ∞

Nº máx. no sistema = K **e Fila máxima** = $K-1$

Tempo de atendimento: Exponencial Negativo

Taxa = μ clientes / u. tempo, servidor

Nº servidores = 1

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

Taxa de ocupação = $\dfrac{\bar{\lambda}}{\mu}$

Taxa de desocupação = $1 - \dfrac{\bar{\lambda}}{\mu}$

$$L_q = \sum_{n=1}^{\infty} (n-1) P_n = L - \frac{\bar{\lambda}}{\mu}$$

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

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

$$W = W_q + \frac{1}{\mu} = \frac{L}{\bar{\lambda}}$$

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}}$$ (taxa de desocupação)

$$P_n = \begin{cases} \rho^n P_0, & \text{se } n = 1,\ldots,K \\ 0, & \text{se } n > K \end{cases}$$

$$P\left(W_q = 0\right) = P_0$$

Características do modelo ($M/M/S(K)$) –> *“fila de espera c/ comprimento limitado e S servidores”*

Chegada: Poisson

Taxa = λ clientes / u. tempo

$$\lambda_n = \begin{cases} \lambda, & \text{se } n = 0,1,\ldots,K-1 \\ 0, & \text{se } n \geq K \end{cases}$$

$$\bar{\lambda} = \sum_{n=0}^{K-1} \lambda P_n = \lambda\left(1 - P_K\right)$$

População = ∞

Nº máx. no sistema = K **e Fila máxima** $= K - S$

Tempo de atendimento: Exponencial Negativo

Taxa = μ clientes / u. tempo, servidor

Nº servidores = S

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

Taxa de ocupação = $\dfrac{\bar{\lambda}}{S\mu}$

Taxa de desocupação = $1 - \dfrac{\bar{\lambda}}{S\mu}$

$$L_q = \sum_{n=S}^{\infty}\left(n - S\right)P_n = \frac{P_0\left(\lambda/\mu\right)^S \rho}{S!\left(1-\rho\right)^2}\left[1 - \rho^{K-S} - \left(K-S\right)\rho^{K-S}\left(1-\rho\right)\right]$$

$$L = \sum_{n=0}^{\infty} nP_n = L_q + \frac{\bar{\lambda}}{\mu}$$

$$W_q = \frac{L_q}{\bar{\lambda}} \qquad W = W_q + \frac{1}{\mu} = \frac{L}{\bar{\lambda}}$$

$$P_0 = 1 \Big/ \left[\sum_{n=0}^{S-1}\frac{\left(\lambda/\mu\right)^n}{n!} + \frac{\left(\lambda/\mu\right)^S}{S!}\sum_{n=S}^{K}\rho^{n-S}\right]$$

$$P\left(W_q = 0\right) = \sum_{n=0}^{S-1} P_n$$

$$P_n = \begin{cases} \dfrac{\left(\lambda/\mu\right)^n}{n!}P_0, & \text{se } n = 1,\ldots,S \\ \dfrac{\left(\lambda/\mu\right)^n}{S!S^{n-S}}P_0, & \text{se } n = S,\ldots,K \\ 0, & \text{se } n > K \end{cases}$$

Características do modelo ($M/M/1/N$) –> *“fila de espera simples, mas população finita”*

Chegada: Poisson

Taxa = λ clientes / u. tempo

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{se } n = 0,1,\ldots,N-1 \\ 0, & \text{se } n \geq N \end{cases}$$

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{N-1} (N-n)\lambda P_n = \lambda(N-L)$$

População = N

Fila máxima = ∞

Tempo de atendimento: Exponencial Negativo

Taxa = μ clientes / u. tempo

Nº servidores = 1

Taxa de ocupação = $\dfrac{\bar{\lambda}}{\mu}$

Taxa de desocupação = $1-\dfrac{\bar{\lambda}}{\mu}$

$$L_q = \sum_{n=1}^{\infty} (n-1) P_n = N - \frac{\lambda+\mu}{\lambda}(1-P_0)$$

$$L = \sum_{n=0}^{\infty} n P_n = L_q + \frac{\bar{\lambda}}{\mu}$$

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

$$W = W_q + \frac{1}{\mu} = \frac{L}{\bar{\lambda}}$$

$$P_0 = 1 \Bigg/ \sum_{n=0}^{N} \left[\frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu} \right)^n \right]$$ (taxa de desocupação)

$$P_n = \begin{cases} \dfrac{N!}{(N-n)!} \left(\dfrac{\lambda}{\mu} \right)^n P_0, & \text{se } n = 1,\ldots,N \\ 0, & \text{se } n > N \end{cases}$$

$$P(W_q = 0) = P_0$$

Características do modelo ($M/M/S/N$) –> "*fila de espera simples c/ S servidores e população finita*"

Chegada: Poisson

Taxa = λ clientes / u. tempo

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{se } n = 0,1,\ldots,N-1 \\ 0, & \text{se } n \geq N \end{cases}$$

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{N-1} (N-n)\lambda P_n = \lambda(N-L)$$

População = N

Fila máxima = ∞

Tempo de atendimento: Exponencial Negativo

Taxa = μ clientes / u. tempo, servidor

Nº servidores = S

Taxa de ocupação = $\dfrac{\bar{\lambda}}{S\mu}$

Taxa de desocupação = $1 - \dfrac{\bar{\lambda}}{S\mu}$

$$L_q = \sum_{n=S}^{\infty} (n-S) P_n$$

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

$$L = \sum_{n=0}^{\infty} n P_n = L_q + \frac{\bar{\lambda}}{\mu}$$

$$W = W_q + \frac{1}{\mu} = \frac{L}{\bar{\lambda}}$$

$$P_0 = 1 \Bigg/ \left[\sum_{n=0}^{S-1} \frac{N!}{(N-n)!\,n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=S}^{N} \frac{N!}{(N-n)!\,S!\,S^{n-S}} \left(\frac{\lambda}{\mu}\right)^n \right]$$

$$P_n = \begin{cases} \dfrac{N!}{(N-n)!\,n!} \left(\dfrac{\lambda}{\mu}\right)^n P_0, & \text{se } n = 1,\ldots,S \\ \dfrac{N!}{(N-n)!\,S!\,S^{n-S}} \left(\dfrac{\lambda}{\mu}\right)^n P_0, & \text{se } n = S,\ldots,N \\ 0, & \text{se } n > N \end{cases}$$

Características do modelo ($M/G/1$) –> *“fila de espera com distribuição genérica”*

Chegada: Poisson

Taxa = λ clientes / u. tempo

População = ∞

Fila máxima = ∞

Tempo de atendimento: qualquer variável aleatória

Média = $\frac{1}{\mu}$

Variância = σ^2

Nº servidores = 1

$\rho = \frac{\lambda}{\mu} < 1$

Taxa de ocupação = ρ

Taxa de desocupação = $1-\rho$

$$L_q = \sum_{n=1}^{\infty} (n-1) P_n = \frac{\lambda^2 \sigma^2 + \rho^2}{2(1-\rho)}$$

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

$$L = L_q + \frac{\lambda}{\mu}$$

$$W = W_q + \frac{1}{\mu} = \frac{L}{\lambda}$$

$P_0 = 1-\rho$ (taxa de desocupação)

Modelo $M/M/1$: Exemplo 2

Um armazém recebe camiões com encomendas que são descarregadas usando empilhadores.

- Os camiões chegam segundo um processo de Poisson à taxa de 16 camiões/dia;
- Os tempos médios de descarga são variáveis (seguindo uma distribuição Exponencial Negativa):

Nº empilhadores	1	2	3	4	5
Tempo médio de descarga (min)	50	20	15	12	10

- A operação de empilhadores custa 7.5 €/h;
- A imobilização dos camiões acarreta um custo de 15 €/h;
- 1 dia = 8 horas.

Pretende-se dimensionar a equipa de empilhadores de modo a minimizar os custos globais do sistema.

Modelo $M/M/1$: Exemplo 2 (resolução)

Taxa de chegada, $\lambda = 2$ camiões/h

Nº de empi-lhadores	Tempo médio de descarga [min]	Taxa de serviço [camiões/h]	Taxa de ocupação	Tempo médio no sistema por camião [h]	Tempo total dos camiões por hora [h]	Custo de imobilização dos camiões [€/h]	Custo dos empi-lhadores [€/h]	Custo total [€/h]
n	$1/\mu$	μ	λ/μ	W	λW	$15\lambda W$	$7.5n$	
1	50	1.20	1.67					
2	20	3.00	0.67	1.00	2.00	30.00	15.00	45.00
3	15	4.00	0.50	0.50	1.00	15.00	22.5	37.50
4	12	5.00	0.40	0.33	0.67	10.00	30.00	40.00
5	10	6.00	0.33	0.25	0.50	7.50	37.50	45.00

Exponencial Negativa? (teste do χ^2 - Qui-Quadrado)

Será razoável admitir que, por exemplo, os intervalos de tempo entre duas chegadas consecutivas (inter-chegadas) seguem aproximadamente uma distribuição Exponencial Negativa?

Suponhamos que se observou a seguinte sequência: $t_1, t_2, ..., t_m$

Uma estimativa razoável para a taxa de chegadas é: $\hat{\lambda} = \dfrac{m}{\sum_{i=1}^{m} t_i}$

Mas será que $t_1, t_2, ..., t_m$ é consistente com a densidade de probabilidades $f(t) = \hat{\lambda} e^{-\hat{\lambda} t}$?

Teste do χ^2 - Qui-Quadrado

Dividindo a amostra em *k* categorias, podemos determinar o número de t_i(s) que caem em cada uma delas: e_j - número esperado, e o_j - número observado.

Então a estatística $\chi^2(obs) = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j}$ deverá seguir uma distribuição χ^2 com (*k*-*r*-1) graus de liberdade (onde *r*=1).

- Se $\chi^2(obs)$ é pequeno, será razoável admitir que os t_i(s) representam uma amostra da distribuição $f(t) = \hat{\lambda} e^{-\hat{\lambda} t}$. (N.B. idealmente todos os $e_j = o_j$).
- Se $\chi^2(obs)$ é grande, esta consideração não será razoável.

Teste do χ^2: (conjeturas)

- Hipótese $H_0 :\ t_1, t_2, ..., t_m$ é uma amostra aleatória da variável aleatória com densidade $f(t)$.
- Hipótese $H_a :\ t_1, t_2, ..., t_m$ não é uma amostra aleatória da variável aleatória com densidade $f(t)$.

Dado um valor de α (erro Tipo I), aceitar H_0 se $\chi^2(obs) \le \chi^2_{k-r-1}(\alpha)$

Exemplo:

(Tempos em minutos.)

Obs.(s)
0.01
0.07
0.03
0.08
0.04
0.1
0.05
0.1
0.11
1.17
1.5
0.93
0.54
0.19
0.22
0.36
0.27
0.46
0.51
0.11
0.56
0.72
0.29
0.04
0.73
T=9.19

m=25 observações.

Sum(*t*)=9.19 minutos. → $\hat{\lambda} = 25/9.19$ =2.72 /minuto.

Consistente com $f(t) = 2.72e^{-2.72t}$?

Escolhendo *k*=5 categorias, a probabilidade de cair em cada uma delas é 0.2, logo $e_j = 0.2 \times 25 = 5$.

Determinam-se as fronteiras das categorias a partir da função acumulada $F(t) = P(A \le t) = 1 - e^{-2.72t}$:

Cat. 1: $0 \le t < l_1$ minutos
Cat. 2: $l_1 \le t < l_2$ minutos
Cat. 3: $l_2 \le t < l_3$ minutos
Cat. 4: $l_3 \le t < l_4$ minutos
Cat. 5: $l_4 \le t$ minutos

$F(l_1) = 0.2$, $F(l_2) = 0.4$, $F(l_3) = 0.6$ e $F(l_4) = 0.8$

Logo tiram-se: $l_1 = 0.08$, $l_2 = 0.19$, $l_3 = 0.34$, $l_4 = 0.59$

.....

Tira-se que $\chi^2(obs) = 0.4$, e escolhendo $\alpha = 0.05$, então $\chi_3^2(0.05) = 7.81$ (tabela).

Conclusão: ACEITAR H_0! (erro 5%)

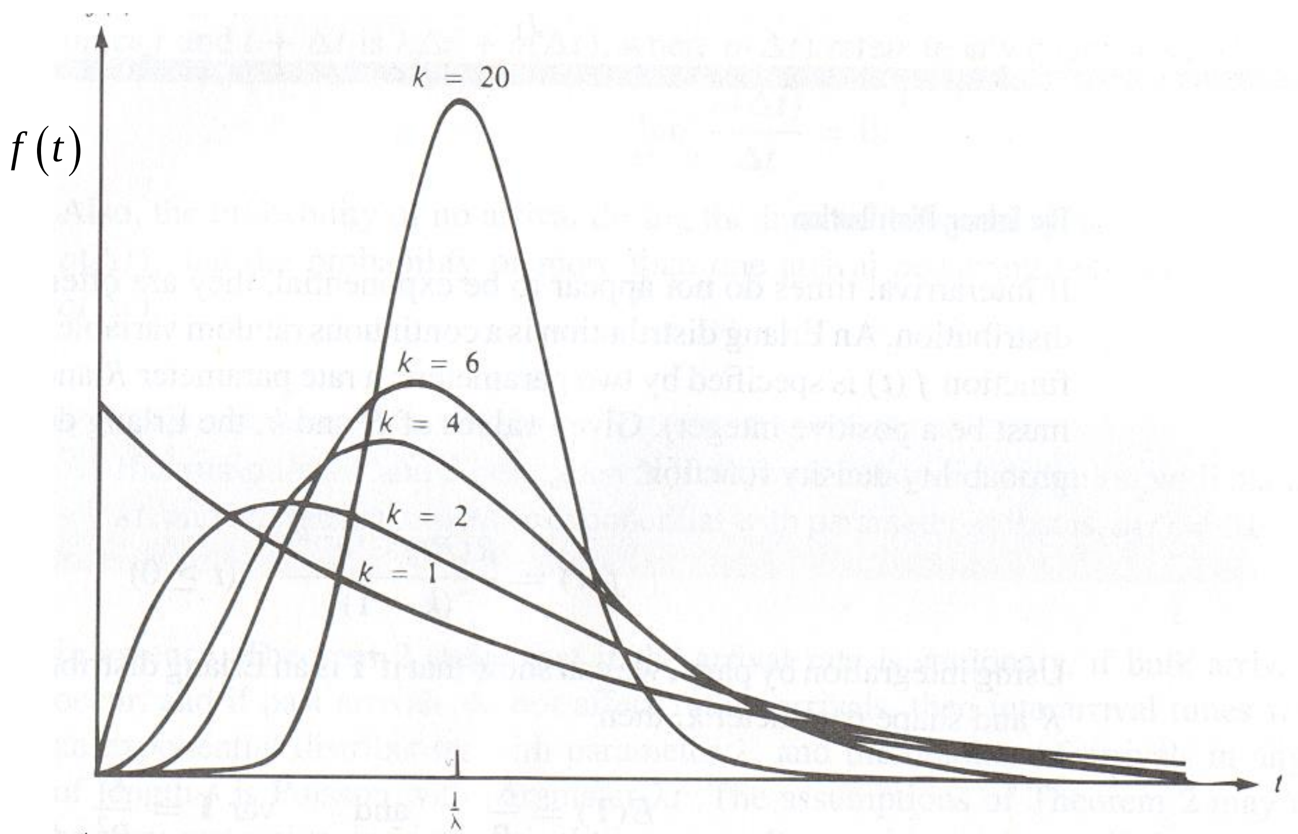
Características da Distribuição *k*-Erlang (E_k)

Função densidade de probabilidade (FDP):

$$f(t) = \begin{cases} \dfrac{\alpha(\alpha t)^{k-1} e^{-\alpha t}}{(k-1)!} & (t \geq 0) \\ 0 & (t < 0) \end{cases}$$

Dois parâmetros: α é a taxa ou escala da distribuição

k é o parâmetro de forma (*k*>0)



Média: $E(T) = \dfrac{k}{\alpha}$ Variância: $Var(T) = \dfrac{k}{\alpha^2}$

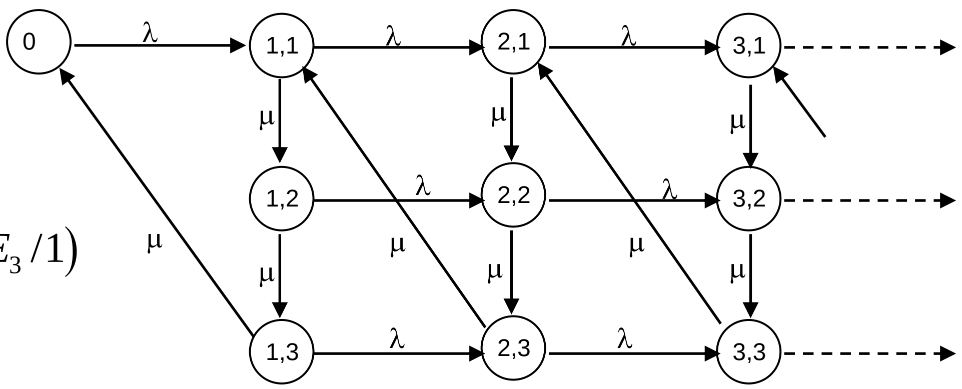
Características da Distribuição *k*-Erlang (E_k) – *cont.*

Dependendo do valor de *k*, a distribuição de Erlang pode adaptar-se a um largo espectro de formas:

- *k*=1, a Erlang coincide com a Exponencial Negativa (assimétrica, "alongada para a direita");
- à medida que *k* aumenta, a distribuição torna-se mais e mais similar à Normal;
- *k* →∞, a Erlang aproxima-se de uma constante (Var = 0)!

Se *X* é uma variável aleatória *IID* segundo uma Erlang $(k, k\alpha)$, então esta pode ser representada pela soma $X = Y_1 + Y_2 + ... + Y_k$, na qual as variáveis Y_i são *IID* Exponenciais Negativas com taxa $(k\alpha)$. Diz-se que o processo (ex., de chegada) de Erlang é equivalente a um processo Exponencial com *k* fases. Exemplo:

$(M / E_3 / 1)$



Modelos Analíticos *vs.* Técnica da Simulação

Os modelos de Filas de Espera que apresentados anteriormente, baseados em distribuições Exponenciais Negativas, foram resolvidos por via analítica. Recorrendo a manipulações matemáticas de reduzida complexidade, foi possível chegar a fórmulas sintéticas para calcular as suas medidas de desempenho fundamentais.

Em muitos casos em que não é razoável admitir uma Exponencial Negativa, pode ser, no entanto, possível admitir outro tipo de distribuição, como por exemplo a Erlang. Porém, a complexidade da via de resolução analítica frequentemente cresce severamente ou torna-se mesmo inviável. (Em alguns casos muito particulares, é possível ainda assim recorrer a tabelas estatísticas para determinar algumas medidas de desempenho em função dos parâmetros admitidos.) Em resumo, apesar da complexidade que as expressões possam aparentar, a abordagem analítica permite obter resultados rápidos e económicos, exatos ou aproximados.

Em última instância, perante um problema concreto que viola fortemente todas as hipóteses dos modelos analíticos disponíveis, recorre-se à técnica da SIMULAÇÃO. A simulação reproduz o funcionamento do sistema, é mais flexível relativamente ao modelo considerado, mas exige geralmente mais recursos computacionais, mais tempo, e um grande cuidado no tratamento estatístico e interpretação dos resultados obtidos.

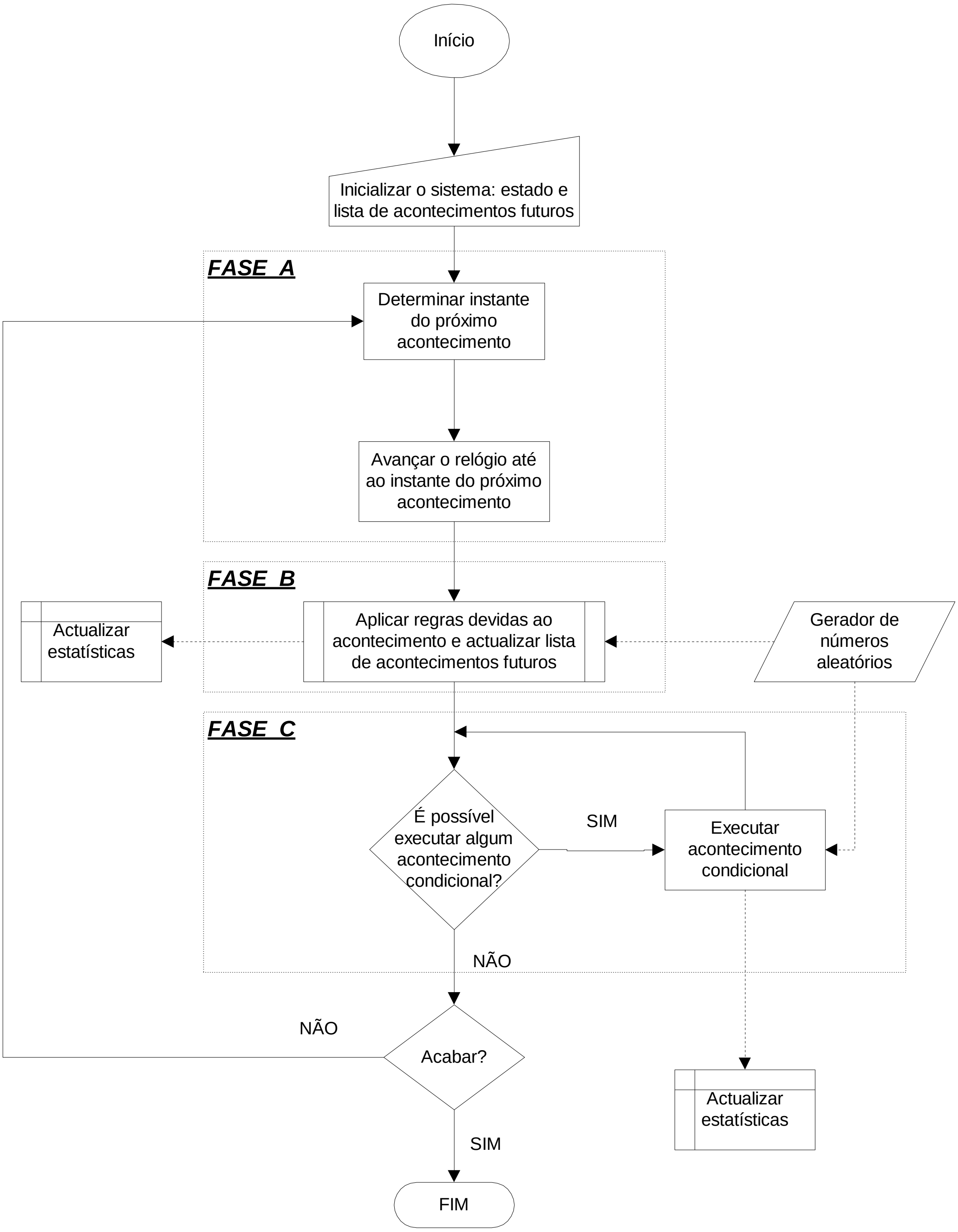
Técnica da Simulação – Elementos principais

Um modelo de simulação "por acontecimentos" inclui os seguintes elementos principais:

- Um relógio (p/ controlo do tempo);
- Um gerador de números aleatórios;
- Entidades;
- Acontecimentos e estados;
- Um conjunto de regras (para cada acontecimento);
- Uma lista atualizada de (instantes p/ futuros) acontecimentos;
- Um processo de coleção e gravação de estatísticas; e
- Um procedimento (fluxograma) que una e dinamize os elementos anteriores, garantindo a execução do processo de funcionamento.

Técnica da Simulação – Procedimento

O fluxograma de três fases (A-B-C) para uma simulação por acontecimentos discretos pode ser descrito da seguinte forma:



Técnica da Simulação – Gerador de aleatoriedade

Normalmente, um gerador de números aleatórios constrói-se a partir de um gerador *standard* de valores contínuos uniformemente distribuídos entre 0 e 1 ($U \in [0,1]$). Por exemplo, no *Excel*^TM temos disponível a função geradora *RAND()* – <u>*ver anexo com outras fórmulas úteis.*</u>

<u>Como obter então um gerador de valores aleatórios (*ex*, tempo *t*) geridos por uma...</u>

<u>...(1) Distribuição Exponencial Negativa?</u>

A partir da Função Cumulativa da distribuição $F(t) = 1 - e^{-\alpha t}$,

resolvendo em ordem a *t*, tira-se que $t = -\frac{1}{\alpha}\ln\left(1 - F(t)\right)$. Como $0 \le F(t) \le 1 \Leftrightarrow 0 \le \left(1 - F(t)\right) \le 1$, e supondo esta quantidade como uma variável aleatória *U* uniformemente distribuída, vem que $\boxed{t = -\frac{1}{\alpha}\ln(U)}$ será uma variável aleatória distribuída segundo uma distribuição Exponencial Negativa com parâmetro α.

<u>...(2) Distribuição *k*-Erlang?</u>

Como a *k*-Erlang é a soma de *k* Exponenciais ($k\alpha$), vem que:

$$t(k, k\alpha) = -\frac{1}{k\alpha}\left[\ln(U_1) + \ln(U_2) + \ldots + \ln(U_k)\right]$$

$$\therefore \qquad \boxed{t(k, k\alpha) = -\frac{1}{k\alpha}\ln\left(\prod_{j=1}^{k} U_j\right)}$$

Simulação discreta numa folha de cálculo (exemplo)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**SIMULADOR M/M/1**			Taxa de chegadas(λ) =		10	/h												
2				Taxa de serviço (μ) =		15	/h												
3																			
4	Acontecimento #		Relógio (t)	δ	q	n	Instantes Futuros		Auxiliares para estatísticas				Comprimento da fila de espera						
5	i	Tipo	(minutos)	#Serviço?	#Fila	#Sistema	Chegada	Partida	t(n+1)-t(n)	$\delta * \Delta t$	$q * \Delta t$	$n * \Delta t$	0		1	2	3	4	5
6	0	Inicialização	0	0	0	0	0	9999	---	---		---							
7	1	Chegada	0	1	0	1	5.8004	1.21404	1.2140377	1.21404	0	1.21404							
8	2	Partida	1.21403768	0	0	0	5.8004	9999	4.5863589	0	0	0							
9	3	Chegada	5.80039653	1	0	1	6.52671	8.7367	0.7263095	0.72631	0	0.72631							
10	4	Chegada	6.52670601	1	1	2	6.85263	8.7367	0.3259222	0.32592	0.32592	0.65184							
11	5	Chegada	6.85262821	1	2	3	8.44028	8.7367	1.587652	1.58765	3.1753	4.76296							
12	6	Chegada	8.44028023	1	3	4	11.7345	8.7367	0.2964217	0.29642	0.88927	1.18569							
13	7	Partida	8.73670195	1	2	3	11.7345	11.1911	2.4543521	2.45435	4.9087	7.36306							
14	8	Partida	11.191054	1	1	2	11.7345	14.5348	0.5434916	0.54349	0.54349	1.08698							
15	9	Chegada	11.7345456	1	2	3	17.1644	14.5348	2.8002518	2.80025	5.6005	8.40076							
16	10	Partida	14.5347974	1	1	2	17.1644	15.6229	1.0881243	1.08812	1.08812	2.17625							
17	11	Partida	15.6229217	1	0	1	17.1644	16.0271	0.4042235	0.40422	0	0.40422							
18	12	Partida	16.0271452	0	0	0	17.1644	9999	1.1372183	0	0	0							

B7 = IF(G6<H6;"Chegada";"Partida")

C7 = IF(B7="Chegada";G6;H6)

F7 = IF(B7="Chegada";F6+1;F6-1)

D7 = IF(F7>0;1;0)

E7 = IF(F7>1;F7-1;0)

G7 = IF(B7="Chegada";C7-log(RAND())/$F$1;G6)

H7 =
| H6 se aconteceu uma "Chegada";
| C7-log(RAND())/$F$2 se aconteceu uma "Partida" e *n*>0, ou se aconteceu uma "Chegada" e H6=9999;
| "9999" se aconteceu uma "Partida" e *n*=0 (i.e, o sistema está vazio).

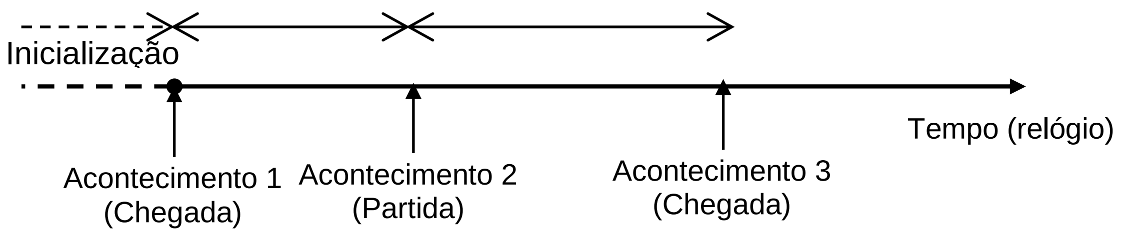
Notas sobre o exemplo de simulação

Entidades: Clientes que usam o sistema M/M/1

Estado (*n*): Nº de clientes no sistema

Acontecimentos: Chegadas e Partidas (ou fins de serviço)

O comportamento geral de um simples *"run"* (fazer F9) do modelo de simulação anterior (ver aplicativo "*io2_simulador_ee1.xls*" – folha de cálculo "*M-M-1 demo*"), pode ser esquematizado da seguinte forma:



Dado que os resultados finais de um único *"run"* podem vir influenciados pelas condições iniciais e pela própria duração da simulação, é conveniente realizar vários *"runs"* com durações razoavelmente longas, e basear as conclusões (sobre as medidas de desempenho do sistema) nos valores médios globais respectivos obtidos ao longo de todo o processo. Para isso, é aconselhável usar as outras folhas de cálculo do mesmo aplicativo.

Qual deverá ser a duração da simulação?

Problema:

Determinar qual a duração da simulação, para que o valor das variáveis do sistema (ex., medidas de desempenho) sejam estimadas com uma determinada acurácia, verificando um dado grau de confiança estatística.

Procedimento normalmente proposto:

1) Executar algumas simulações curtas, utilizando diferentes séries de números aleatórios (e.g. partindo de diferentes sementes);

2) Estimar a média e a variância da variável que se pretende medir (*ex.*, o comprimento médio da fila, o tempo médio de permanência no sistema,…), e admitir para esta uma lei de distribuição Normal;

3) Sabendo que o desvio padrão dessa variável é inversamente proporcional à raiz quadrada da duração da simulação, calcular que duração deve ter uma experiência de simulação, para obter uma dada acurácia a um dado grau de confiança estatística.

Qual deverá ser a duração da simulação? (exemplo)

Valor médio
(obtido com curtas simulações de 10 horas) = *120* unidades
Desvio padrão = *40* unidades

Qual a duração mínima de uma simulação, para que o nível médio de inventário de *120* possa ser especificado com uma acurácia de $\pm$ *5* unidades, com uma confiança estatística de *90%*?

Resposta:

Admitindo uma lei Normal: $Z = \dfrac{X-\mu}{\sigma} \;\rightarrow\; \sigma_D = \dfrac{X-\mu}{Z_{90\%}}$

tira-se que o desvio padrão a obter deve ser de:

$$\sigma_D = \frac{5}{1.6449} = 3.04 \text{ unidades.}$$

Fazendo então a proporcionalidade inversa: $\dfrac{\sigma_D}{40} = \sqrt{\dfrac{10}{D}}$

Obtém-se que a simulação deve corresponder a:

$$D = \frac{10 * 40^2}{(3.04)^2} = 1732 \text{ horas (ou mais).}$$

Qual é a acurácia obtida com a simulação? (exemplo)

No caso do exemplo anterior, se só fosse possível executar uma simulação de duração equivalente a *1000* horas, qual seria a acurácia da estimativa, com um grau de confiança de *90%*?

Resposta:

O desvio padrão seria, proporcionalmente, de $\sigma_D = 40\sqrt{\frac{10}{1000}}$ = *4* unidades.

Logo, a acurácia conseguida na estimativa do valor médio da variável, com uma confiança estatística de *90%*, seria de $\pm$*4 * 1.6449*, ou seja de aproximadamente $\pm$ *6.6* unidades.

ANEXO: Algumas funções de interesse no Excel^{TM}

Random Generation:	RAND()	0 to 1 random generator
	RAND()*(b-a)+a	a to b random generator
	e.g. RAND()*100	0 to100 random generator

N.B. Introducing RAND(), and then pressing F9, instead of ENTER, transforms formula into value.

Normal Distribution: NORMDIST(x; mean; std; cumulative)

Cumulative if cumulative=TRUE, mass if FALSE

e.g. NORMDIST(1.65; 0; 1; TRUE) = 0.9505

NORMINV(prob; mean; std), the cumulative inverse

e.g. NORMINV(0.95; 0; 1) = 1.6449

Conditional: IF(logical_test; value_if_true; value_if_false)

N.B. Up to 7 IFs can be nested.

Logical IS functions: ISBLANK(value), ISTEXT(value), ISERROR(value)

e.g. ISERROR(AVERAGE(B1:B5))

Maths/Stats: ABS, AVERAGE, LOG, POWER, ROUND, STDEV, SQRT, SUM, SUMIF

Tools → Goal Seek… → Set cell … To value … By changing cell …

Adjusts the value of a cell (independent variable), so that the other cell (dependent variable or formula) equals the seek value.