FACULDADE DE ECONOMIA
Universidade de Coimbra

RELATÓRIO DE UMA AULA TEÓRICO-PRÁTICA

Teoria das Filas de Espera

Provas de Aptidão Pedagógica e Capacidade Científica

Luis Miguel Cândido Dias

COIMBRA 1995

ÍNDICE

1 - ÂMBITO DO TRABALHO

O tema escolhido para a aula teórico-prática a que se refere este relatório consta actualmente do programa de duas disciplinas de duas licenciaturas diferentes: Investigação Operacional (licenciatura em Economia) e Métodos de Apoio à Decisão (licenciatura em Organização e Gestão de Empresas). Optou-se por tratar o tema no contexto do programa de Investigação Operacional (licenciatura em Economia), dado que o autor é responsável por algumas turmas práticas dessa disciplina no ano lectivo 1994/95. Não haveria, no entanto, alterações significativas caso se optasse por inserir o tema no contexto da disciplina de Métodos de Apoio à Decisão. A carga horária semanal da disciplina de Investigação Operacional (licenciatura em Economia) é de duas horas teóricas e duas horas práticas.

O programa da disciplina de Investigação Operacional (4º Ano, 2º Semestre, licenciatura em Economia) é o seguinte:

1. Introdução à teoria da decisão
2. Programação linear
3. Algoritmos para problemas de transporte e afectação
4. Optimização em redes
5. Filas de espera

O programa actual carece de dois outros temas com alguma importância para o tema das filas de espera: cadeias de Markov e simulação. O tratamento pedagógico do tema da teoria das filas de espera é influenciado pela ausência desses temas (principalmente o primeiro) do programa.

Assumir-se-á que os alunos estão familiarizados com conceitos de análise matemática e estatística leccionados em disciplinas que antecedem Investigação Operacional no plano da licenciatura em Economia.

O tema da teoria das filas de espera, embora seja o último item do programa, é importante no âmbito do programa da disciplina, dado que ilustra uma parte significativa da investigação operacional: a dos modelos probabilísticos. É ainda importante pelo facto de ser um dos métodos de investigação operacional mais leccionados nas universidades e aplicados nas organizações [Hillier e Lieberman, 1990]; [Lane et al., 1993]. Por fim, atendendo a que os futuros economistas poderão exercer funções de gestão ou consultoria, é possível que o tema seja por estes aplicado a problemas da vida profissional.

A teoria das filas de espera nasceu de trabalhos na área das telecomunicações no início do século. Posteriormente, desenvolveu-se a par das restantes técnicas de investigação operacional e tem sido aplicada em áreas muito distintas nos sectores secundário e terciário. Nas publicações

científicas de investigação operacional pode encontrar-se evidência de pesquisa activa a nível metodológico em filas de espera, assim como aplicações bem sucedidas em indústrias, instituições bancárias e instituições de saúde, para nomear apenas alguns exemplos.

Para preparar a presente aula houve necessidade de proceder a uma recolha de material respeitante ao tema. Essa recolha não foi de modo algum exaustiva, circunscrevendo-se apenas ao que o autor considerou pertinente para uma aula a futuros licenciados em Economia. Dessa recolha resultou um texto, que se apresenta em apêndice, que contém um resumo de aspectos que se consideram interessantes para ensinar aos alunos da licenciatura em Economia. A preocupação básica na selecção efectuada foi a de manter uma "perspectiva de Gestor", com ênfase na modelação, o reflectiu as expectativas que podem recair sobre a acção de apoio à gestão numa organização:

- Identificar situações que podem ser modeladas por sistemas de filas de espera;
- Construir um ou vários modelos de compromisso, suficientemente realistas para o propósito do gestor, mas ao mesmo tempo suficientemente simples tendo em vista a obtenção de resultados;
- Se o modelo for suficientemente simples, obter soluções analíticas para o modelo;
  se não, ter cultura específica para pesquisar textos avançados e para dialogar com especialistas.

O planeamento da aula, cuja base é o resumo apresentado em apêndice, teve de atender à incerteza sobre o tempo lectivo disponível para leccionar o tema. Dado que o número de horas lectivas de um semestre tem sido sistematicamente inferior ao planeado e ao desejável, o tema da teoria das filas de espera, último tema do programa, nunca foi leccionado. Por esse motivo, apesar de se pensar que duas semanas de aulas sejam adequadas ao desenvolvimento do tema, é provável que o tempo disponível se reduza apenas a uma, ou mesmo a nenhuma.

O presente relatório (excluindo o apêndice) parte do princípio que só está disponível uma aula para leccionar o tema e, embora beneficiasse da existência de uma aula teórica prévia, não o pressupõe. Sendo de duas horas a duração de uma aula, será impossível apresentá-la na sua totalidade durante a prestação da Prova.

Optou-se por incluir o apêndice neste relatório de modo a expor de modo implícito algumas opções tomadas e os conteúdos que seriam aprofundados no caso de haver tempo para uma segunda aula. Como consequência da inclusão do apêndice, haveria demasiada redundância caso se apresentasse o planeamento da aula sob a forma habitual de um desenvolvimento textual. Em vez disso, a aula é apresentada de forma esquemática, explicitando objectivos, conteúdos e a estratégia a seguir. Embora seja de leitura mais difícil, este tipo de abordagem é o habitualmente eleito no ensino de pedagogia e não enferma do pressuposto da existência de um texto a ser lido (eventualmente após memorização) na aula.

2 - PLANEAMENTO E RESUMO DA AULA

O apêndice apresenta um desenvolvimento metodológico da teoria das filas de espera, podendo constituir simultaneamente a base de um texto de apoio aos alunos da disciplina. O desenvolvimento apresentado no apêndice, bem como o planeamento da aula, resultaram sobretudo da convicção de que não se deve sobreavaliar a importância do cálculo de resultados para um dado modelo. Considera-se importante evitar que dado modelo seja utilizado sem que estejam percebidos os pressupostos do mesmo como delimitadores da sua adequação à realidade. Por outras palavras, pretende-se evitar que os alunos obtenham resultados correctos para um modelo inadequado. Nessa perspectiva achou-se interessante incluir no desenvolvimento em apêndice uma secção como a 5.3, de carácter quase anedótico, mas que apresenta aos alunos o caminho por vezes esquecido que distancia a teoria da aplicação prática. Esta secção está incluída no capítulo 5, problemas de decisão, ao qual se atribui grande importância, que deriva de ser a decisão o fim último da teoria das filas de espera.

A aula teórico-prática a que se refere o presente relatório está planeada para um tempo lectivo de duas horas. No entanto, é de admitir que essa duração seja insuficiente, sobretudo se os alunos tiverem a receptividade esperada e participarem activamente na aula ou, pelo contrário, revelarem dificuldades em acompanhar o ritmo proposto. Sendo a temática das filas de espera o último item do programa da disciplina de Investigação Operacional, o número de horas disponível para o apresentar é muito incerto e provavelmente escasso. Por esse motivo, apresentam-se os conteúdos por uma ordem que parte do geral para o particular e do imprescindível para o acessório.

A importâcia atribuída à situação de decisão, bem como o seu carácter potencialmente motivador, conduziu à sua escolha para início da aula. Apela-se à intuição dos alunos para que, ainda que desconhecendo os conteúdos a leccionar, compreendam a situação básica de compromisso entre os custos associados à espera em filas e os custos associados à redução dessa espera.

A aula continua com a exposição do modelo de fila de espera e suas características, culminando na notação de Kendall. De seguida apresentam-se as medidas de desempenho no estado estacionário, insistindo-se no carácter estocástico do processo que se pretende descrever e na natureza do estado estacionário. Exemplifica-se a problemática da escolha das medidas de desempenho através do descrito em [Kolesar, 1984]. A fórmula de Little será mostrada sob a forma que [Ravindran, 1987] aconselha — $L = R\ W$ — em vez da habitual $L = \lambda\ W$, mais restritiva. Para esta, como para as restantes relações existentes no estado estacionário, apela-se à intuição do aluno para que este aceite os resultados sem recurso a demonstrações formais. Terminar-se-á esta parte com um exemplo.

Após ter contactado com uma visão global dos sistemas e modelos de filas de espera, o aluno será desafiado a ponderar algumas questões, a que se dará resposta no decorrer da aula:

- Será suficiente ter uma taxa de serviço superior à taxa de chegada para se proporcionar um bom nível de serviço?
- Se a taxa de serviço de um servidor A for duas vezes superior à taxa de serviço de um servidor B, pode-se concluir que os tempos de espera médios num sistema com o primeiro servidor serão metade dos de um sistema com o segundo?
- Ao ter dois servidores em vez de um, o tempo de espera é reduzido em metade?

Nesta altura entra-se nos conteúdos mais técnicos que constituem o tema dos modelos de Markov. Em primeiro lugar, tenta-se munir o aluno de conhecimentos que lhe permitam discernir se é ou não justificável a adopção da distribuição exponencial nos seus modelos. Seguidamente parte-se dos gráficos dos processos nascimento-morte, com breve referência aos seus pressupostos e às equações de Kolmogorov, para obter soluções analíticas. Esta escolha constitui um compromisso entre abordagens escolhidas por autores que exploram com alguma profundidade as cadeias de Markov (p. ex. [Hillier e Lieberman, 1990]), e abordagens encontradas noutros autores que apresentam apenas os resultados sem qualquer justificação para os mesmos (p. ex. [Anderson et. al., 1991]). O objectivo é exemplificar como se obtêm soluções analíticas para alguns modelos e tornar o aluno apto a calcular soluções para outros modelos de Markov. Omite-se a utilização de tabelas e gráficos para obtenção de soluções, por se considerar que as calculadoras e os computadores têm relegado tais elementos auxiliares para segundo plano.

Após a apresentação dos processos de nascimento e morte propõe-se aos alunos um exemplo de aplicação concreta, seguido do exemplo do modelo M/M/1 genérico. De seguida, representam-se os diagramas de transição para alguns modelos genéricos mais complexos como exercício.

A aula conclui-se com exemplos de tomada de decisão utilizando modelos de Markov. Estes exemplos permitem responder às questões levantadas anteriormente.

O pouco tempo disponível implica que temas como o factor humano, os modelos não markovianos, a obtenção de resultados e as redes de filas de espera fiquem de fora de uma primeira aula. Acentua-se porém que uma segunda aula onde estes temas fossem abordados seria de inegável interesse.

3 - AULA: OBJECTIVOS, CONTEÚDOS E ESTRATÉGIAS

3.1. Motivação

Objectivos:

- Conhecer o propósito da teoria das filas de espera;
- Reconhecer na teoria das filas de espera um método da investigação operacional probabilístico e descritivo;
- Compreender o compromisso envolvido na situação de decisão.

Conteúdos:

- Propósito da teoria das filas de espera;
- Natureza da teoria das filas de espera;
- Problemática de decisão.

Estratégias:

a) Exposição oral focando os seguintes aspectos:

• Dar exemplos de formação de filas de espera como um fenómeno comum, especialmente os não associados ao sentido comum do vocábulo "fila".

• Incluir o exemplo de uma viagem por via aérea. Primeiro poderá ser necessário esperar numa fila para comprar a passagem. O vendedor do bilhete terá de verificar se há lugar no avião, eventualmente esperando numa fila de espera pela resposta. No terminal de embarque forma-se novamente uma fila, assim como para entrar no avião. Ao avião é atribuído um número que indica a sua vez na fila de espera para descolar, e quando finalmente chega ao destino tem novamente de esperar por uma pista de aterragem e por um terminal de desembarque. O passageiro terá então de esperar numa fila para desembarcar e terá que esperar pela sua bagagem. Conclui deste modo a sua viagem aérea e, pacientemente, chega à fila de espera para apanhar um taxi...

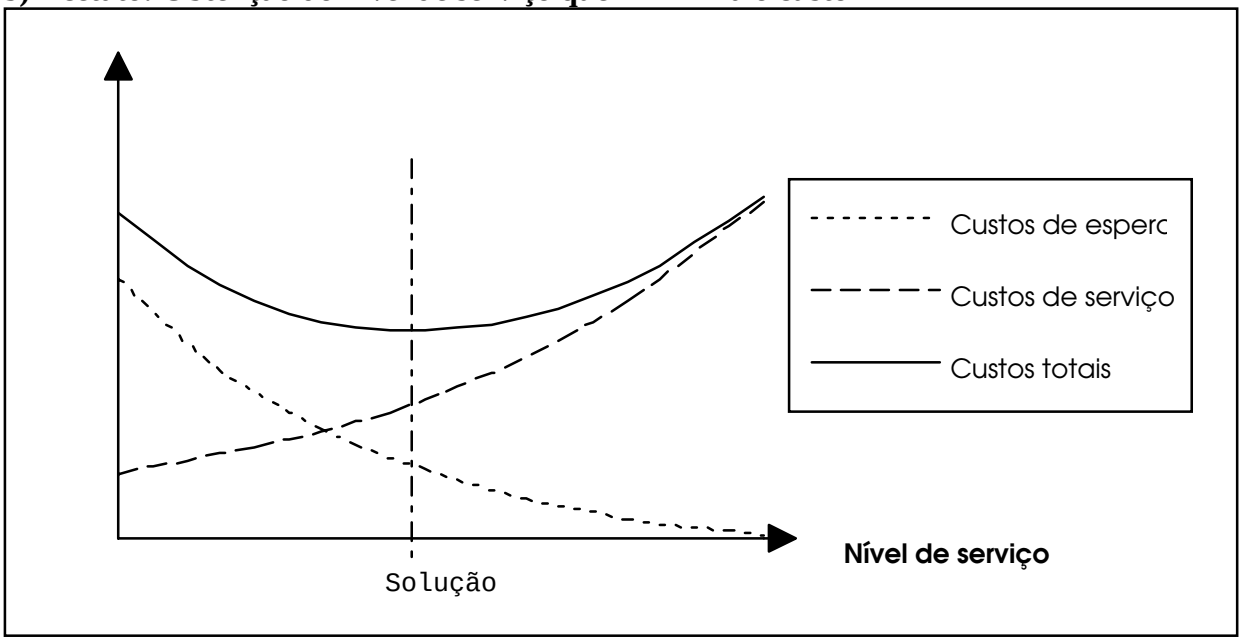
• A análise de filas de espera é um método descritivo e probabilístico.

• O objectivo da teoria das filas de espera (também designada por teoria de sistemas estocásticos de serviço) é o de obter modelos adequados de situações em que se formem filas, de modo a prever o seu comportamento. Esse comportamento, expresso por diversas medidas de desempenho, pode servir de base à tomada de decisões sobre o sistema em causa.

• A situação de decisão: escolha entre sistemas alternativos anteriormente à sua colocação em funcionamento.

• A tomada de decisão: compromisso entre os custos associados à espera em filas e os custos da redução dessa espera.

b) Acetato: Obtenção do nível de serviço que minimiza o custo



b') Exposição oral focando os seguintes aspectos:

• o custo total esperado é igual ao custo esperado de fornecer um determinado nível de serviço, adicionado ao custo esperado pela estada em fila resultante do nível de serviço oferecido.

• Exemplos de custos: salário de fucionários que prestam o serviço (custo de serviço), perda de bens voláteis, tempo não produtivo de funcionários e má imagem com que o cliente fica (custos de espera).

3.2. O modelo de fila de espera

Objectivos:

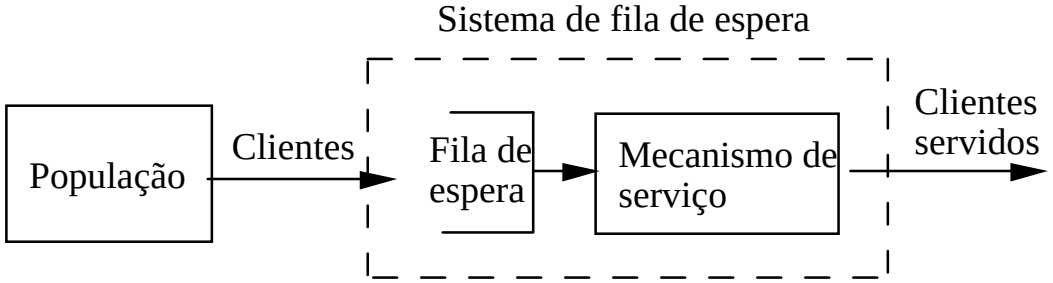
- Indicar os componentes básicos de um modelo de fila de espera;
- Identificar componentes básicos dada uma situação apropriada a um modelo de fila de espera;
- Descrever as características de um modelo de fila de espera;
- Classificar filas de espera através da notação de Kendall.

Conteúdos:

- Componentes básicos de um modelo de fila de espera;
- Características de um modelo de fila de espera;
- Notação de Kendall.

Estratégias:

a) Acetato: Modelo de fila de espera



a') Exposição oral focando os seguintes aspectos:

• Os componentes de um modelo de fila de espera são a população ou fonte de potenciais clientes e o sistema de fila de espera (que pode designar-se apenas por "sistema"). O sistema é constituído pela fila de espera propriamente dita e pelo mecanismo de serviço que serve os elementos na fila de espera. O mecanismo de serviço, que pode compreender um ou mais servidores, serve os clientes por uma ordem determinada pela disciplina de serviço.

• Um cliente é uma entidade contável originária de determinada população, que espera pela sua vez de ser servido na fila de espera (considera-se então que já entrou no sistema), ocupa um servidor durante um certo tempo e por fim deixa o sistema.

• Exemplos: chegada de aviões a um aeroporto, conjunto de máquinas que avaria frequentemente requerendo o serviço de um reparador, central telefónica.

• Identificação de uma unidade cliente e de uma unidade servidora, através de exemplos (uma família que compra bilhetes para o teatro; um conjunto de mecânicos que reabastece de combustível um automóvel de fórmula 1) e de uma regra prática (número máximo de clientes que é possível servir simultaneamente).

b) Acetato:

Para se construir um modelo de fila de espera há que tomar em conta os seguintes factores:
- natureza da população;
- lei estatística que descreve o número de chegadas (de clientes) por unidade de tempo;
- capacidade do sistema;
- disciplina de serviço;
- mecanismo de serviço;
- lei estatística que descreve a duração do serviço prestado a cada cliente;

b') Exposição oral focando os seguintes aspectos:

(População)

• Uma das características mais importantes da população é o seu tamanho (finito ou infinito).

• A análise do sistema torna-se geralmente mais simples quando se considera uma população infinita. Na prática, as situações em que há um conjunto muito numeroso de clientes potenciais podem ser modeladas como tal. Por exemplo, a população que utiliza o bar da Faculdade de Economia pode ser considerada infinita, embora em rigor não o seja.

• Uma regra prática é a de considerar a população infinita em todos os casos, excepto aqueles em que o número de clientes que pode chegar ao sistema num dado intervalo de tempo dependa significativamente do tamanho da população nesse intervalo. Por exemplo, numa situação com 10 máquinas e um reparador a probabilidade de se avariar uma máquina na próxima hora depende do número de máquinas que esteja a funcionar.

(Processo de chegada)

• Processo determinístico ou estocástico. Num caso como no outro interessa sobretudo a taxa de chegada, que é o número de clientes que em média chega ao sistema por unidade de tempo.

• Nesta aula considera-se que os clientes chegam independentemente uns dos outros e individualmente.

(Capacidade do sistema)

• Define-se capacidade como o número máximo de clientes que pode estar simultaneamente no sistema de fila de espera (pode considerar-se infinita).

• A introdução de uma capacidade finita no modelo implica maior complexidade.

• Num sistema com capacidade finita admite-se a hipótese de um cliente não entrar no sistema por este ter a capacidade esgotada. Assume-se então que o cliente não tenta regressar.

• Exemplos: salas com servidores ATM e servidores ATM na via pública.

(Disciplina de serviço)

• A disciplina de serviço é uma regra (ou um conjunto de regras) que define a ordem pela qual os clientes são servidos.

• Disciplinas mais utilizadas:

- ordem de chegada (FCFS), p. ex. barbeiro;
- disciplina de pilha (LCFS), p. ex. processamento de uma pilha de documentos;
- ordem aleatória (RSS), p. ex. processamento de peças todas iguais;
- pré-escalonada, p. ex. consultório médico;
- esquemas com prioridades
    - com preempção, p. ex. serviço de urgências do hospital X;
    - sem preempção, p. ex. serviço de cirurgia do hospital Y.

• Por omissão, assume-se que a disciplina consiste em servir os clientes por ordem de chegada.

(Mecanismo de serviço e servidores)

• Necessidade de determinar o número de servidores, a sequência pela qual são visitados por cada cliente e o processo de serviço de cada servidor descrito pela lei estatística (ou determinística, i.e. um número fixo) do tempo de serviço a cada cliente.

• Quando nada é indicado em contrário, assume-se que os servidores operam em paralelo e os tempos de serviço são variáveis aleatórias independentes e identicamente distribuídas.

• Um parâmetro importante é a taxa de serviço, que é o número médio de clientes servidos por unidade de tempo, por cada servidor.

(Outras características)

• Comportamentos de "balking" (recusa), "reneging" (desistência) e "jockeying" (apostas).

• Custos associados aos comportamentos anteriores (exemplos: oportunidade de negócio perdida, abandono de compras em supermercados).

• Compromisso entre realismo do modelo e viabilidade de obter resultados.

c) Quadro: Notação de Kendall

processo de chegada / processo de serviço / número de servidores/...
Os códigos utilizados são:
M (de Markov) para tempos entre chegadas (tempos de serviço) exponencialmente distribuídos
D para tempos entre chegadas (tempos de serviço) determinísticos
E_k para tempos entre chegadas (tempos de serviço) seguindo uma lei Erlang-k
G (de genérico) para tempos entre chegadas (tempos de serviço) seguindo uma lei qualquer

c') Exposição oral focando os seguintes aspectos:

• Notação de Kendall como veículo de comunicação.

• Modelos genéricos (p. ex. M/M/s).

• Extensão da notação: o quarto símbolo representa a capacidade do sistema (por omissão infinita), o quinto representa o tamanho da população (por omissão infinita) e o sexto representa a disciplina de serviço de acordo com os acrónimos ingleses (por omissão FCFS).

c'') Exercício:

Indicar o significado de M/M/s/K, M/M/s/s, M/D/1/K/K e M/M/4/10/10/RSS.

3.3. Medidas de desempenho

Objectivos:

- Empregar a notação matemática usada para quantificar sistemas de fila de espera;
- Seleccionar em cada situação as medidas de desempenho mais interessantes;
- Distinguir estado transitório de estado estacionário;
- Conhecer as relações básicas que existem no estado estacionário;
- Reconhecer o interesse das relações existentes no estado estacionário.

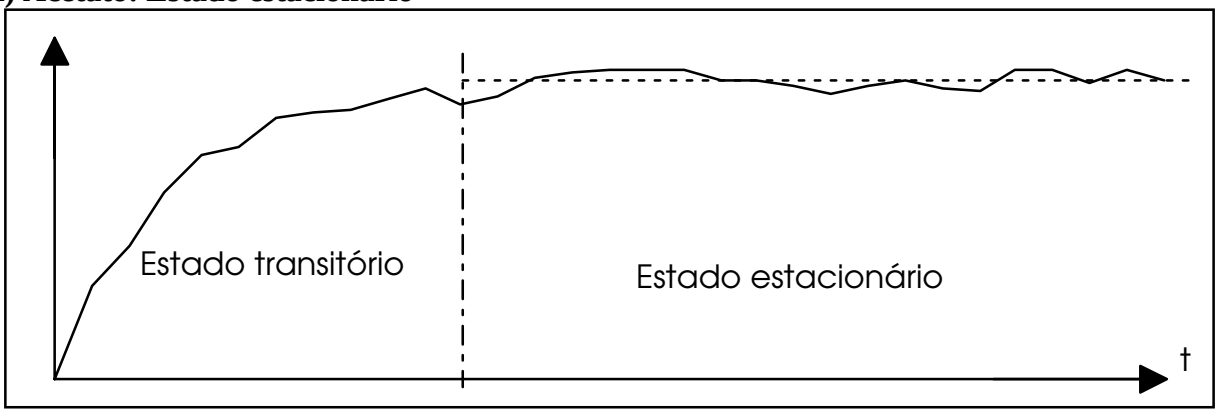
Conteúdos:

- Notação matemática;
- Estado transitório e estado estacionário;

- Medidas de desempenho;
- Relações básicas que existem no estado estacionário.

Estratégias:

a) Acetato: Estado estacionário



a') Exposição oral focando os seguintes aspectos:

• No início do seu funcionamento o sistema atravessa um estado transitório que poderá evoluir para um estado estacionário.

• Assume-se nesta aula que se verificam as condições suficientes para que o sistema atinja o estado estacionário.

• A principal característica deste estado não é um comportamento regular, mas sim um comportamento em que as medidas de desempenho não dependem de há quanto tempo o sistema está em funcionamento nem do estado inicial do sistema.

b) Acetato: Pressupostos

- a população é um conjunto discreto de clientes, finito ou infinito;
- os clientes chegam ao sistema a uma taxa fixa;
- qualquer cliente que entre no sistema é servido;
- os servidores de um sistema têm todos a mesma taxa de serviço;
- a taxa de serviço (por servidor) não varia com o estado do sistema.

c) Acetato: Notação

Especificação do modelo:

λ taxa de chegada dos clientes ($1/\lambda$ é o tempo médio entre chegadas)

μ taxa de serviço de um servidor ($1/\mu$ é o tempo médio de serviço a um cliente)

s número de servidores

ρ factor de utilização do sistema

Medidas de desempenho:

p_i Probabilidade de estarem i clientes no sistema

b_i	Probabilidade de estarem ocupados i servidores
q_i	Probabilidade de estarem i clientes na fila de espera
L	Número médio de clientes no sistema
L_q	Número médio de clientes na fila de espera
B	Número médio de servidores ocupados
R	Taxa de saída média, ou taxa de chegadas efectivas (entradas)
W	Tempo médio que um cliente está no sistema
W_q	Tempo médio que um cliente está na fila de espera

c') Exposição oral focando os seguintes aspectos:

• Entre as medidas de desempenho, as letras minúsculas indicam probabilidades, enquanto as letras maiúsculas indicam esperanças matemáticas.

• R indica a taxa de saída de clientes do sistema, que é igual à taxa de chegadas efectiva , dado que para um sistema esteja no estado estacionário se verifica que o número médio de entradas é igual ao número médio de saídas.

• O símbolo λ pode ter uma interpretação distinta em modelos com população finita.

• Não há medidas universalmente mais importantes.

• Geralmente procura-se uma medida de desempenho que espelhe os custos associados à formação de filas de espera.

• Exemplo exposto em [Kolesar, 1984]

d) Acetato: Relações básicas

$$L = \sum_{i=0}^{\infty} i\, p_i \;,\; L_q = \sum_{i=0}^{\infty} i\, q_i = \sum_{i=s}^{\infty} (i-s)\, p_i \quad \text{e} \quad B = \sum_{i=0}^{s-1} i\, p_i + \sum_{i=s}^{\infty} s\, p_i$$

$L = L_q + B$ (porque B é o número médio de clientes a ser atendido)

$W = W_q + (1/\mu)$ (porque $1/\mu$ é o tempo médio de serviço a um cliente)

$\rho = B / s$ (porque pode definir-se ρ deste modo)

Capacidade e população infinitas => $R = \lambda$ e $B = \lambda / \mu$

Capacidade finita (K) => $R = \lambda (1 - p_K)$

Lei de Little $L = R\, W$ $(L_q = R\, W_q)$

d') Exposição oral focando os seguintes aspectos:

• Explicar, tão brevemente quanto possível, cada relação omitindo demonstrações.

• Descrever o procedimento habitual de começar pela determinação da lei da variável aleatória discreta que indica o número de clientes no sistema no estado estacionário.

e) Exemplo

Apresenta-se agora um exemplo (adaptado do original apresentado em [Jensen, 1986]) que ilustra a aplicação das relações existentes no estado estacionário.

Uma estação de correios possui três postos de atendimento. Ocasionalmente formam-se filas de espera (sistema de fila única). Efectuou-se um estudo no qual foram recolhidos os seguintes dados:

tempo médio entre chegadas ($1/\lambda$) = 30 segundos (0,5 minutos);

tempo médio de atendimento ($1/\mu$) = 1,25 minutos;

número médio de clientes no sistema (L) = 6 clientes.

Apenas com estes dados é já possível indicar bastantes características deste sistema. Na falta de mais informação assume-se um modelo G/G/3. Algumas medidas podem ser obtidas:

taxa de chegada: $\lambda = 1/0,5 = 2$ (clientes por minuto);

taxa de serviço: $\mu = 1/1,25 = 0,8$ (clientes por minuto);

tempo médio de espera no sistema: $W = L / R = L / \lambda = 6 / 2 = 3$ minutos;

tempo médio de espera na fila: $W_q = W - (1/\mu) = 3 - 1,25 = 1,75$ minutos;

nº médio de clientes na fila de espera: $L_q = R\ W_q = \lambda\ W_q = 2 \times 1,75 = 3,5$ clientes;

utilização: $\rho = B / s = \lambda / s\mu = 2/(3 \times 0,8) = 0,8(3)$.

Este exemplo mostra o interesse da teoria das filas de espera ao evidenciar uma aparente contradição. De facto, a utilização é de aproximadamente 83%, o que implica que, se o trabalho for igualmente dividido entre os servidores, cada servidor não tem nada que fazer durante 17% do tempo (cerca de dez minutos por cada hora). No entanto, em média, estão 3,5 clientes na fila de espera. A natureza estocástica do processo pode conduzir a que durante algum tempo não esteja nenhum cliente na fila de espera, e que noutra altura esteja um grande número de clientes (maior que 3,5) aguardando ser atendido.

f) Propõem-se as seguintes questões para reflexão:

- Será suficiente ter uma taxa de serviço superior à taxa de chegada para se proporcionar um bom nível de serviço?
- Se a taxa de serviço de um servidor A for duas vezes superior à taxa de serviço de um servidor B, pode-se concluir que os tempos de espera médios num sistema com o primeiro servidor serão metade dos de um sistema com o segundo?
- Ao ter dois servidores em vez de um o tempo de espera é reduzido em metade?

3.4. Modelos de Markov

Objectivos:

- Saber a importância dos processos de Poisson na teoria das filas de espera e suas implicações na modelação e resolução de problemas;

- Compreender como se obtêm soluções analíticas para alguns sistemas simples.

Conteúdos:

- Propriedades da lei exponencial;
- Modelos de Markov (caracterização e resolução);
- Processos nascimento-morte (diagramas de transição de estado).

Estratégias:

a) Exposição oral focando os seguintes aspectos:

• Necessidade de calcular parâmetros através de soluções analíticas ou estimar os parâmetros através de simulação.

• Modelos de Markov como um tipo de modelo para o qual é possível obter soluções analíticas.

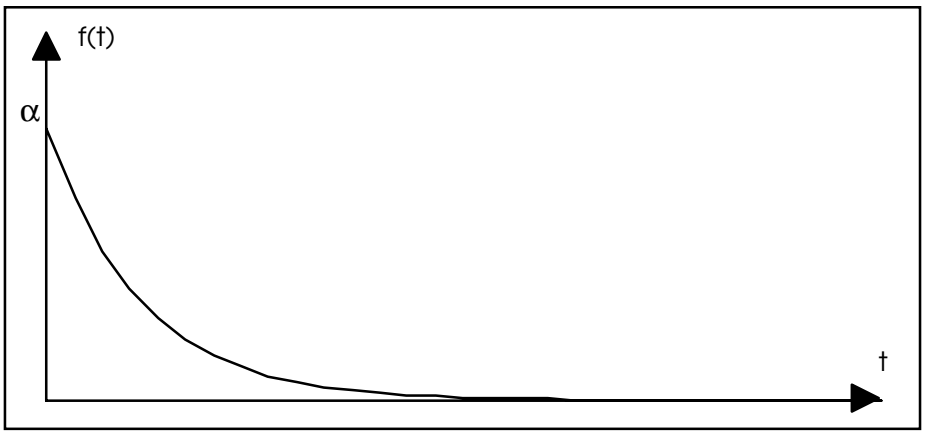
• Processos de Poisson como característica dos modelos de Markov.

• Problema da adequação dos modelos de Markov à realidade.

b) Acetato: Distribuição exponencial negativa

A Função densidade de probabilidade (f.d.p.) de uma variável aleatória contínua T que segue uma lei exponencial de parâmetro α é:

$$f_T(t) = \begin{cases} \alpha\, e^{-\alpha t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$



Algumas propriedades importantes:

1) f(t) é uma função decrescente em sentido estrito;
2) ausência de memória, ou seja, P(T > t + x | T > x) = P(T > t);
3) sejam n v.a.r. independentes $T_1 \sim e(\alpha_1)$, $T_2 \sim e(\alpha_2)$, ... , $T_n \sim e(\alpha_n)$; então
   $U = \min \{ T_1, T_2, \ldots, T_n \} \sim e(\alpha)$, com $\alpha = \_ \alpha_i$;

4) as leis exponencial e de Poisson de parâmetro α estão relacionadas, referindo-se a primeira ao tempo entre dois acontecimentos consecutivos e a segunda ao número de acontecimentos por unidade de tempo.

b') Exposição oral focando os seguintes aspectos:

• Histograma, elaborado a partir de dados colhidos por amostragem dos tempos entre chegadas (ou tempos de serviço), como meio de falsificação.

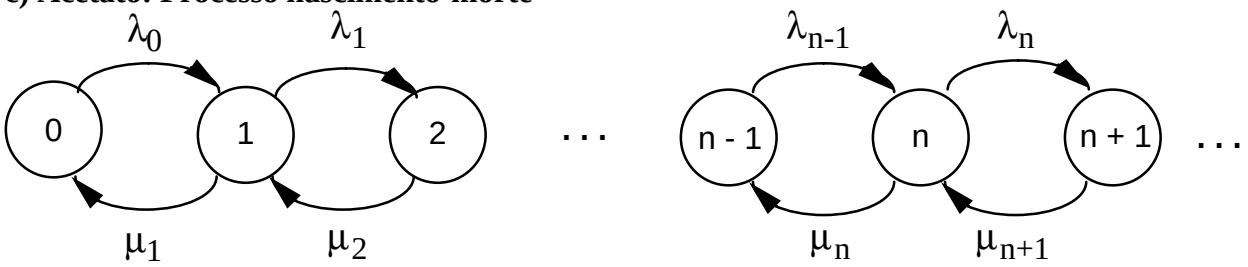
• Justificação da adopção da lei exponencial para a distribuição dos tempos entre chegadas.

• Possibilidade de utilizar uma lei exponencial para modelar a chegada de clientes de vários tipos, ignorando as distinções, desde que cada tipo de cliente chegue de acordo com leis exponenciais.

• Justificação da adopção da lei exponencial para a distribuição dos tempos de serviço.

• Modelos de Markov como uma aproximação à realidade.

c) Acetato: Processo nascimento-morte



c') Exposição oral focando os seguintes aspectos:

• Processos de "nascimento e morte" como um meio para obtenção de soluções analíticas para modelos de Markov.

• Diagramas de transições.

• Equações "de balanço" e sua resolução.

d) Exemplo

Um modelo com poucos estados [Jensen, 1986]

Considere-se um barbeiro cuja barbearia têm três cadeiras: uma para servir os clientes e as restantes para os clientes esperarem. Os clientes chegam à barbearia de acordo com um processo de Poisson com um taxa média de chegada de 4 por hora. O tempo de serviço a cada cliente segue uma distribuição exponencial com média de 10 minutos. O sistema tem capacidade finita e 4 estados, representados no diagrama de transições da figura seguinte. O número de clientes é decrementado sempre que um dos clientes é servido (o que sucede com taxa μ=6 por hora) e é incrementado quando um cliente entra no sistema.

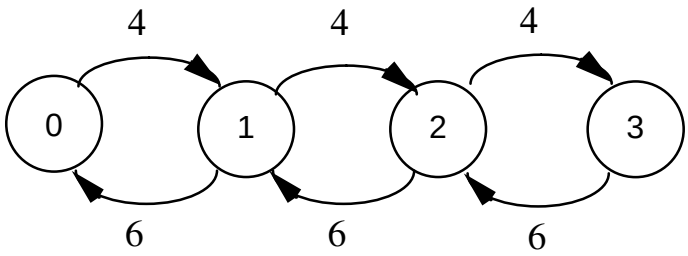


Diagrama de transições para um barbeiro e duas cadeiras de espera (M/M/1/3)

De acordo com as expressões já conhecidas resulta para o estado estacionário:

$C_1 = 4 / 6 = 0{,}667$; $C_2 = 4^2 / 6^2 = 0{,}444$; $C_3 = 4^3 / 6^3 = 0{,}296$

$p_0 = 1 / (1 + 0{,}667 + 0{,}444 + 0{,}296) = 0{,}415$

$p_1 = 0{,}667\ p_0 = 0{,}277$; $p_2 = 0{,}444\ p_0 = 0{,}185$; $p_3 = 0{,}296\ p_0 = 0{,}123$.

Outras medidas pretendidas podem agora ser calculadas, por exemplo:

número médio de clientes no sistema — $L = 0\ p_0 + 1\ p_1 + 2\ p_2 + 3\ p_3 = 1{,}016$ clientes;

taxa média de entradas (saídas) — $R = \lambda\ (1 - p_3) = 4\ (1 - 0{,}123) = 3{,}508$ clientes/hora;

tempo médio de espera no sistema — $W = L / R = 0{,}290$ horas (cerca de 17 minutos);

número médio de clientes na fila — $L_q = 1\ p_2 + 2\ p_3 = 0{,}431$ clientes;

tempo médio na fila de espera — $W_q = L_q / R = 0{,}123$ horas (cerca de 7 minutos);

utilização — $\rho = B = L - L_q = 0{,}585$ (58,5%).

A percentagem de clientes que se perde por falta de capacidade é 12,3% (p_3), apesar de o barbeiro só estar ocupado pouco mais de metade do tempo (B).

Se de uma das cadeiras para espera fosse transformada numa cadeira de serviço e se contratasse mais um barbeiro, o processo seria representado pela figura seguinte. Quando há mais que um cliente no sistema ambos os barbeiros estarão a trabalhar e a taxa de serviço conjunta duplica, passando a 12 clientes por hora.

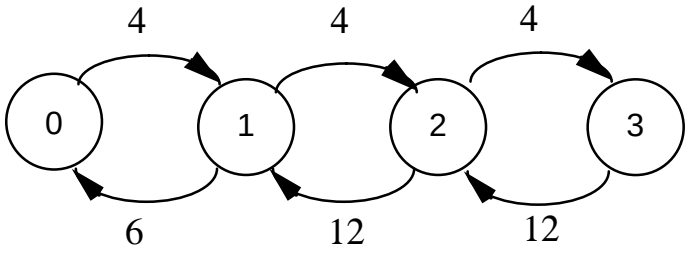


Diagrama de transições para dois barbeiros e uma cadeira de espera (M/M/2/3)

Novamente torna-se simples encontrar as medidas pretendidas. Alguns resultados são:

percentagem de clientes que se perde — $p_3 = 0{,}038$ (3,8%);

tempo médio na fila de espera — $W_q = 0{,}0098$ horas (0,58 minutos).

Cabe agora ao barbeiro decidir se está disposto a pagar a um colega e fazer a alteração na cadeira em troca da melhoria do desempenho do sistema.

e) Exemplo

Modelo M/M/1

A figura seguinte apresenta o diagrama de transições para um modelo de Markov com um servidor, capacidade infinita e população infinita. A análise do modelo é simplificada pelo facto de as taxas médias de chegada λ e de serviço μ serem constantes.

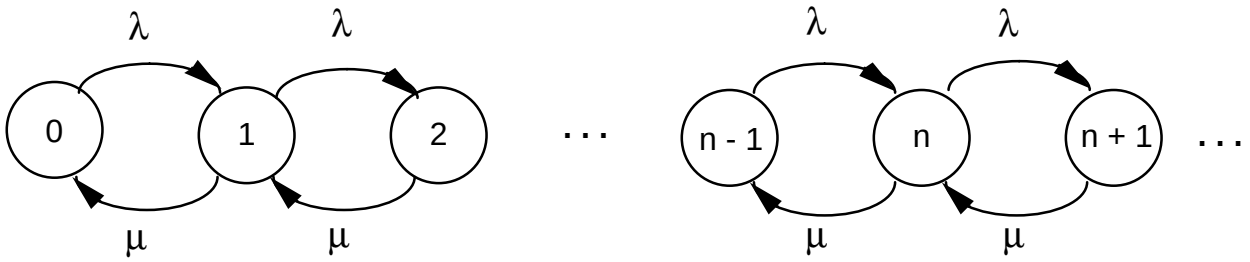


Diagrama de transições de um modelo M/M/1

A partir da solução do sistema de equações de balanço e das relações existentes no estado estacionário obtém-se:

$C_i = (\lambda / \mu)^i = \rho^i$, para i = 1, 2, ... (dado que, para um sistema uniservidor, $\rho = \lambda / \mu$)

$p_i = \rho^i p_0$, para i = 1, 2, ...

$$p_0 = \frac{1}{1+\sum_{i=1}^{\infty}\rho^i} = \frac{1}{\sum_{i=0}^{\infty}\rho^i} = \left(\frac{1}{1-\rho}\right)^{-1} = 1-\rho$$

o que permite escrever $p_i = (1 - \rho) \rho^i$, para i = 1, 2, ...

$$L = \sum_{i=0}^{\infty} i(1-\rho)\rho^i = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

$$L_q = \sum_{i=1}^{\infty}(i-1)p_i = \sum_{i=1}^{\infty} i p_i - \sum_{i=1}^{\infty} p_i = L - (1-p_0) = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

A fórmula de Little permitiria então obter W e W_q. O uso todos estes resultados pressupõe que $\mu > \lambda$, caso contrário a fila de espera cresceria sem limite.

Um resultado talvez inesperado, que resulta de $p_i = (1 - \rho) \rho^i$, é o de que, para sistemas M/M/1 capazes de servir todas as chegadas ($\rho<1$), o estado mais provável do sistema é aquele em que o servidor está desocupado e não há clientes na fila!

f) Exercício: (Ver solução na secção 4.2 do apêndice)
Desenhar os diagramas de transições para os modelos M/M/s, M/M/s/K e M/M/s/N/N.

3.5. Situações de decisão

Objectivo:

- Utilizar as medidas de desempenho na tomada de decisões;

Estratégias:

a) Exercício:

Considere um sistema em que um empregado de escritório verifica determinado tipo de embalagens. Pretende-se limitar o número de embalagens na fila de espera a 5. Supondo que a situação é modelada por um modelo M/M/1:

i) qual será a percentagem de tempo que o empregado está inactivo?

[$L_q = \rho^2 / (1-\rho)$ -> $1-\rho$ 0,15 -> 15%]

ii) caso a percentagem de tempo inactivo fosse considerada excessiva e se aumentasse a taxa a que chegam as embalagens de modo a reduzir essa percentagem para 5%, qual seria o número médio de embalagens na fila de espera? [$L_q = \rho^2 / (1-\rho)$ -> L_q 18]

Discussão: Este exemplo permite responder à primeira questão colocada no final da secção 3.3. O modelo M/M/1 revela o inevitável conflito entre a maximização da utilização do sistema e a manutenção de filas de espera não muito longas. A ideia de que é suficiente proporcionar uma taxa de serviço μ igual à taxa de chegadas λ está neste caso errada. O valor ideal para ρ é inferior a 1, dependendo, para cada aplicação específica, dos custos da ineficiência dos servidores relativamente ao tamanho das filas de espera.

b) Exemplo:

1ª Parte

Uma empresa é constituída por dois departamentos que pretendem conjuntamente alugar uma fotocopiadora para criar um pequeno centro de cópias. A escolha deve ser feita entre dois modelos alternativos: a fotocopiadora F100 consegue duplicar em média 20 trabalhos por hora e custa Esc. 5000 por dia, enquanto a fotocopiadora F200 duplica 21 trabalhos por hora com um custo de Esc. 7000 por dia. O centro estará aberto 10 horas por dia e tem uma procura média de 18 trabalhos por hora. Os funcionários são originários dos dois departamentos e efectuam eles próprios as suas duplicações. A administração considera que cada hora que um funcionário permanece inactivo neste sistema de fila de espera acarreta uma perda de lucro de Esc. 500.

Considerando que este exemplo cumpria os pressupostos do modelo M/M/1, ter-se-ia:

	Fotocopiadora F100	Fotocopiadora F200
Taxa de chegadas λ	18 trabalhos/h	18 trabalhos/h
Taxa de serviço μ	20 trabalhos/h	21 trabalhos/h
Utilização ρ	0,9 (90%)	0,857 (85,7%)
Tempo de espera W	0,5 h	0,33(3) h

Custo da espera	18 x 0,5 x 500 = Esc. 4500/h	18 x 0,33(3) x 500 = Esc. 3000/h
Custo total	10 x 4500 + 5000 = Esc. 50000/dia	10 x 3000 + 7000 = Esc. 37000/dia

A escolha da administração recaiu, com base nestes resultados, na fotocopiadora F200. Chama-se a atenção para o facto de uma pequena alteração na taxa de serviço (um aumento de 5%) conduzir a uma redução significativa dos tempos de espera (cerca de 33%). Este exemplo mostra como estão errados raciocínios que considerem que aumentar a taxa de serviço equivale a reduzir na mesma proporção os tempos de espera médios. Destarte se conclui que é negativa a resposta à segunda questão colocada no final da secção 3.3.

2ª Parte: A administração, através do estudo efectuado, verificou que os custos de espera eram bastante superiores aos custos da oferta do serviço. Por esse motivo, resolveu efectuar um outro estudo, que considerasse o aluguer de uma segunda fotocopiadora F200, de modo a reduzir ainda mais os tempos de espera. Duas hipóteses foram estudadas: atribuir uma fotocopiadora a cada departamento ou partilhar as duas fotocopiadoras pelos dois departamentos. A primeira situação pode ser descrita por dois sistemas M/M/1, enquanto a segunda conduz a um modelo M/M/2. Considerando que os trabalhos a duplicar provinham de ambos os departamentos a igual taxa (9 trabalhos por hora) os resultados obtidos foram:

	Uma só fotocopiadora	Uma fotocopiadora para cada departamento	Duas fotocopiadoras partilhadas
Taxa de chegadas λ	18 trabalhos/h	9 trabalhos/h (*)	18 trabalhos/h
Taxa de serviço μ	21 trabalhos/h	21 trabalhos/h (*)	21 trabalhos/h (*)
Utilização ρ	0,857 (85,7%)	0,429 (42,9%)	0,429 (42,9%)
Tempo de espera W	0,33(3) h (20 min.)	0,083 h (5 min.)	0,058 (3,5 min)
Custo da espera	18 x 0,33(3) x 500 = Esc. 3000/h	2 x (9 x 0,083 x 500) = Esc. 750/h	18 x 0,058 x 500 = Esc. 525/h
Custo total	10 x 3000 + 7000 = Esc. 37000/dia	10 x 750 + 2 x 7000 = Esc. 21500/dia	10 x 525 + 2 x 7000 = Esc. 19250/dia

(*) por servidor

Este exemplo permite retirar uma conclusão que pode contrariar o senso comum: ao duplicar o número de servidores (modelo M/M/2) o tempo de espera médio foi reduzido em 82,5%, e não em metade, o que responde à terceira questão colocada no final da secção 3.3. Uma outra conclusão, já de acordo com a intuição, é a de que um modelo M/M/2 (dois servidores com fila comum) proporciona melhor desempenho do que dois modelos M/M/1 independentes (dois servidores, cada um com sua fila). Esta última conclusão é válida para outros modelos, mesmo

não markovianos, pelo que é cada vez mais frequente encontrar sistemas com vários servidores e fila única.

APÊNDICE - PROBLEMAS DE FILAS DE ESPERA

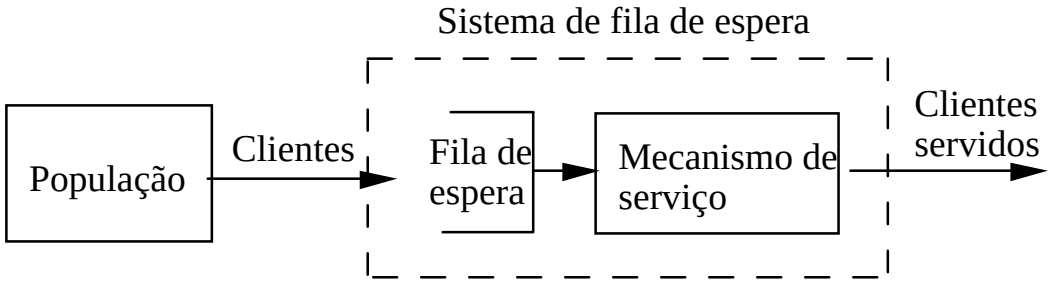
1. Introdução

A formação de filas de espera é um fenómeno comum. Ocorre frequentemente em repartições públicas, bancos, aeroportos, supermercados e outros. Para além destes exemplos associados ao sentido comum do vocábulo "fila" existem outros, menos óbvios, que exibem comportamentos análogos. Disso são exemplo materiais *à espera* de ser processados numa fábrica, máquinas *à espera* de reparação, ou pacientes aguardando a vez de serem consultados.

A análise de filas de espera é um método descritivo, porque descreve um processo, em oposição a métodos de optimização (normativos) que visam encontrar a solução óptima de determinado problema. É também um método probabilístico, uma vez que estuda um fenómeno aleatório[1].

O objectivo da teoria das filas de espera[2] é o de obter modelos adequados de situações em que se formem filas, de modo a prever o seu comportamento. Esse comportamento, expresso por diversas medidas de desempenho, pode servir de base à tomada de decisões sobre o sistema em causa. A tomada de decisão consiste frequentemente num compromisso entre os custos de um menor desempenho do sistema e os custos de construir um sistema mais oneroso.

Para uma dada situação em que ocorram filas, a teoria das filas de espera opera sobre um modelo. Os componentes de um modelo de fila de espera são a **população** ou **fonte** de potenciais **clientes** e o **sistema** de fila de espera (doravante designado apenas por "sistema"). O sistema é constituído pela **fila de espera** propriamente dita e pelo **mecanismo de serviço** que serve os elementos na fila de espera (figura A1). O mecanismo de serviço, que pode compreender um ou mais **servidores**, serve os clientes por uma ordem determinada pela **disciplina de serviço**.



[1] Mais concretamente, a teoria das filas de espera lida com um processo estocástico de estados discretos e tempo contínuo. As mudanças de estado podem ocorrer a qualquer instante, e o número de estados é finito ou infinito numerável.

[2] Por vezes encontram-se as designações "teoria de sistemas estocásticos de serviço", "teoria da congestão" e "teoria do teletráfego", dado que muitos sistemas de interesse não permitem a formação de filas de espera (sistemas com perda). No entanto, as duas últimas designações podem dar a entender um campo de aplicação limitado.

Figura A1 - Modelo de fila de espera

Tanto os clientes como os servidores podem não ser humanos. No âmbito deste texto considera-se que um cliente é uma entidade contável originária de determinada população, que espera pela sua vez de ser servido na fila de espera (considera-se então que já entrou no sistema), ocupa um servidor durante um certo tempo e por fim deixa o sistema. Por exemplo, ao modelar-se a chegada de aviões a um aeroporto como uma fila de espera considerar-se-ia que a população seria o universo de aviões que pode aterrar no aeroporto, os clientes seriam os aviões, e os servidores seriam as diversas pistas de aterragem. Um outro exemplo, mais surpreendente, é o do conhecido problema de existir um conjunto de máquinas que avaria frequentemente requerendo o serviço de um reparador. Estes casos podem modelar-se como filas de espera em que a população é a totalidade de máquinas a funcionar, os clientes na fila de espera são as máquinas avariadas, e os servidores são os reparadores existentes. Uma fila de espera pode portanto existir sem que os clientes estejam fisicamente "em fila".

2. O modelo de fila de espera

Para se construir um modelo de fila de espera há que tomar em conta os seguintes factores:

- natureza da população;
- lei estatística que descreve o número de chegadas (de clientes) por unidade de tempo;
- capacidade do sistema;
- disciplina de serviço;
- mecanismo de serviço;
- lei estatística que descreve a duração do serviço prestado a cada cliente;

Em muitas aplicações da teoria das filas de espera alguns destes factores não permanecem constantes no tempo. Nestes casos há que construir um modelo para cada período de tempo em que o factor se pode considerar constante.

2.1. População

Uma das características mais importantes da população é o seu tamanho. Interessa sobretudo saber se é considerada de tamanho finito ou infinito. A análise do sistema torna-se geralmente mais simples quando se considera uma população infinita. Na prática, as situações em que há um conjunto muito numeroso de clientes potenciais podem ser modeladas como tal.

Uma regra possível para optar entre um modelo e outro é a de considerar a população infinita em todos os casos, excepto aqueles em que o número esperado de solicitações (chegadas) de clientes dependa significativamente do tamanho da população. O tamanho da população será, em cada instante, igual ao tamanho do universo dos clientes subtraindo aqueles que estão no sistema.

2.2. Processo de chegada

Considera-se que clientes podem chegar ao sistema de um modo determinístico (p. ex. um cliente chega de exactamente 10 em 10 minutos) ou de um modo estocástico de acordo com uma lei estatística. Tanto num caso como no outro interessa sobretudo a taxa de chegada, que é o número de clientes que em média chega ao sistema por unidade de tempo. A unidade de tempo (p. ex. minuto, hora, dia, ...) pode ser escolhida convenientemente de acordo com o problema.

Neste texto considera-se que os clientes chegam independentemente uns dos outros e individualmente. Há no entanto modelos em que os clientes chegam em grupos, que têm merecido a atenção de investigadores. Só se considera que chegam clientes em grupo se estes forem servidos individualmente.

2.3. Capacidade do sistema

Define-se capacidade como o número máximo de clientes que pode estar simultaneamente no sistema de fila de espera. Recorde-se que um cliente está no sistema se estiver na fila de espera ou se estiver a ser servido. Geralmente assume-se que a capacidade é infinita, o que implica que a fila de espera pode crescer sem limite. A introdução de uma capacidade finita no modelo implica maior complexidade, pelo que só se deve faze-lo quando se está convicto de que tal é necessário para modelar convenientemente o funcionamento do sistema. Num sistema com capacidade finita admite-se a hipótese de um cliente não entrar no sistema por este ter a capacidade esgotada. Assume-se então que o cliente se perde e procura serviço noutro local.

Imagine-se por exemplo uma estação de serviço que tem espaço para servir um veículo e para conter três outros veículos em fila de espera. Este sistema terá capacidade de quatro caso seja proibido esperar fora da estação de serviço (na via pública) e terá capacidade infinita caso contrário. Um outro exemplo [Kolesar, 1984] é o de um banco com uma pequena divisão contendo alguns servidores ATM (caixas automáticas). Nessa pequena divisão cabem doze pessoas (a utilizar os ATM ou à espera). No caso de os clientes não formarem filas de espera fora da divisão quando esta estiver cheia então o sistema deve ser modelado como sendo de capacidade finita. Caso contrário, considerar uma capacidade infinita é o mais apropriado, tal como uma situação em que os ATM estivessem na via pública.

2.4. Disciplina de serviço

A disciplina de serviço é uma regra (ou um conjunto de regras) que define a ordem pela qual os clientes são servidos. Habitualmente considera-se que são servidos por ordem de

chegada (FCFS, First Come-First Served, na literatura em língua inglesa), sobretudo se os clientes forem humanos pois considera-se que é a ordem socialmente justa. Caso nada se indique em contrário é esta a disciplina de serviço considerada.

Outras disciplinas de serviço comuns são:

- disciplina de pilha (LCFS, Last Come-First Served, na literatura em língua inglesa): o último cliente a chegar é o primeiro a ser servido;
- ordem aleatória (RSS, Random Service Selection, na literatura em língua inglesa): o próximo cliente a ser servido é escolhido ao acaso;
- pré-escalonada: os clientes são servidos de acordo com uma ordem pré-estabelecida.

Para além destas disciplinas de serviço existem outras que envolvem prioridades. Nestas, os clientes são servidos de acordo com o seu nível de prioridade, sendo usada uma das disciplinas do parágrafo anterior para distinguir a ordem pela qual são servidos clientes com a mesma prioridade. Os sistemas com prioridades podem ainda dividir-se em preemptivos e não preemptivos. Um sistema preemptivo implica que um servidor interrompa o serviço prestado a um cliente para se dedicar a outro cliente com maior prioridade entretanto chegado. Nesse caso, podem ainda distinguir-se sistemas em que o serviço de um cliente pouco prioritário interrompido é continuado de sistemas em que o serviço é recomeçado. Num sistema não preemptivo o serviço não é interrompido quando chegam clientes mais prioritários. Os sistemas com prioridades não preemptivos podem ainda incluir um tempo de latência, adiando propositadamente o serviço a clientes não prioritários (mesmo que haja servidores inactivos), de modo a servir mais rapidamente clientes prioritários que possam chegar posteriormente.

2.5. Mecanismo de serviço e servidores

A modelação do mecanismo de serviço determina o número de servidores, a sequência pela qual são visitados por cada cliente e o processo de serviço de cada servidor, descrito pela lei estatística (ou determinística) do tempo de serviço a cada cliente. Quando nada é indicado em contrário, assume-se que os servidores operam em paralelo. Tal significa que um cliente é servido por um dos servidores que esteja livre e depois deixa o sistema. Assume-se igualmente que os tempos de serviço são constantes ou variáveis aleatórias independentes e identicamente distribuídas para todos os servidores.

Um parâmetro importante é a taxa de serviço, que é o número médio de clientes servidos por cada servidor, por unidade de tempo.

Neste texto não se consideram modelos que incluem servidores que “vão de férias" ou avariam, servidores que só entram em funcionamento (com ou sem custos de arranque) quando existem r’ ou mais clientes na fila de espera e desligam quando há menos de r (com r ² r’) e servidores que servem grupos de clientes (p. ex. elevadores ou autocarros).

2.6. Outras características

Há ainda características não mencionadas anteriormente que se podem incorporar em modelos de filas de espera. Geralmente, quanto mais rico é o modelo mais este se aproxima da realidade. Contudo, torna-se mais difícil a obtenção de soluções para o modelo à medida que se adicionam detalhes.

Fenómenos muito interessantes ocorrem quando os clientes são humanos, pois há que contar com o comportamento dos mesmos. O mais comum é o de o cliente se recusar a colocar-se na fila de espera[3] (entrar no sistema) quando esta estiver demasiado longa. Nestes casos o cliente procura outro sistema ou regressa noutra altura. Mais difícil de detectar é o caso de o cliente não entrar no sistema em determinada altura por ter aprendido a evitar o serviço a essas horas. Disso é exemplo um cliente que não se desloque a uma agência bancária a determinada hora por prever que a(s) fila(s) de espera o demorem. Para muitas organizações, os casos em que o cliente acaba por não entrar no sistema reflectem-se em oportunidades de negócio perdidas. Uma situação parecida é a de um cliente desistir do serviço após ter estado algum tempo na fila de espera[4], por antecipar que o tempo que lhe falta esperar é ainda longo. Para além do negócio que se perde, outros custos podem surgir nestes casos. Por exemplo, se um cliente deixar a fila de uma caixa de um supermercado então abandonará por certo as suas compras, eventualmente contendo bens congelados ou igualmente perecíveis.

Por fim há fenómenos de difícil modelação em organizações com múltiplos servidores. No caso de haver uma fila de espera por servidor pode haver clientes que saiam da fila onde entraram e re-entrem para uma fila vizinha[5]. No caso de haver uma fila única existe a possibilidade de um cliente ceder vários lugares na fila de espera para poder ser atendido pelo seu servidor preferido. Tal sucede por vezes em agências bancárias em que se desenvolve alguma empatia entre o cliente e um servidor em particular. Este tipo de fenómenos pode ser ignorado ou não, dependendo dos fins para os quais o modelo está a ser construído.

2.7. Notação de Kendall

Através do que acima foi descrito consegue imaginar-se uma grande variedade de modelos de filas de espera. Para facilitar a discussão destes modelos foi desenvolvido um código para classificar estes modelos, a chamada notação de Kendall.

Na sua forma mais simples, a notação de Kendall consiste em três símbolos separados por uma barra. O formato é:

processo de chegada / processo de serviço / número de servidores

Os códigos utilizados são:

M (de Markov) para tempos entre chegadas (tempos de serviço) exponencialmente distribuídos

3 Na literatura em língua inglesa usa-se o vocábulo "balking" para designar este comportamento.

4 "Reneging" em inglês.

5 "Jockeying" em inglês.

D para tempos entre chegadas (tempos de serviço) determinísticos

E_k para tempos entre chegadas (tempos de serviço) seguindo uma lei Erlang-k

G (de genérico) para tempos entre chegadas (tempos de serviço) seguindo uma lei qualquer

A notação utilizada para sistemas com múltiplos servidores (para apresentar resultados genéricos) usa a letra s (de servidor), como em M/M/s ou M/G/s. Encontra-se também frequentemente a letra c (de canal) por terem sido os canais telefónicos a primeira aplicação da teoria das filas de espera.

A notação pode incluir ainda mais símbolos. O quarto símbolo representa a capacidade do sistema (por omissão infinita), o quinto representa o tamanho da população (por omissão infinita), e o sexto representa a disciplina de serviço de acordo com os acrónimos ingleses apresentados em 2.4. (por omissão FCFS).

O primeiro modelo a ser estudado foi o M/M/s/s no início deste século por A. K. Erlang para a companhia de telefones de Copenhaga. É um exemplo de modelo de fila de espera curioso porque não tem fila de espera. O modelo também é útil para empresas que, não sendo operadoras de serviços de telecomunicações, atendam clientes por via telefónica.

3. Medidas de desempenho

Tendo em conta que o tempo perdido em filas de espera pode constituir um custo para as organizações, entende-se a preocupação dos gestores em melhorar as características dos sistemas de fila de espera considerando diversas alternativas. Para isso é útil quantificar o desempenho de cada sistema alternativo através de medidas de desempenho que constituirão uma valiosa informação para quem decide.

Considera-se que, no início do seu funcionamento, o sistema atravessa um estado transitório que poderá evoluir para um estado estacionário. Assume-se neste texto que se verificam as condições necessárias para que o sistema atinja o estado estacionário. A principal característica deste estado não é um comportamento regular, mas sim um comportamento em que as medidas de desempenho não dependem de há quanto tempo o sistema está em funcionamento nem do estado inicial do sistema. A notação seguinte refere-se a um sistema no estado estacionário.

3.1. Notação utilizada

A literatura sobre a teoria das filas de espera é quase unânime na adopção de uma notação para a quantificação de sistemas de fila de espera. Por esse motivo adopta-se essa notação, apesar de remeter para a língua inglesa. Pressupõe-se que:

- a população é um conjunto discreto de clientes, finito ou infinito;
- os clientes chegam ao sistema a uma taxa fixa;

- qualquer cliente que entre no sistema é servido;
- existe um ou vários servidores em paralelo, todos com a mesma lei de serviço;
- a taxa de serviço por servidor não varia com o estado do sistema.

Usar-se-ão os seguintes símbolos:

Especificação do modelo:

λ	taxa de chegada dos clientes ($1/\lambda$ é o tempo médio entre chegadas)
μ	taxa de serviço de um servidor ($1/\mu$ é o tempo médio de serviço a um cliente)
s	número de servidores
ρ	factor de utilização do sistema

Medidas de desempenho:

p_i	Probabilidade de estarem i clientes no sistema
b_i	Probabilidade de estarem ocupados i servidores
q_i	Probabilidade de estarem i clientes na fila de espera
L	Número médio de clientes no sistema
L_q	Número médio de clientes na fila de espera
B	Número médio de servidores ocupados
R	Taxa de saída média, ou taxa de chegadas efectivas
W	Tempo médio que um cliente está no sistema
W_q	Tempo médio que um cliente está na fila de espera

Entre as medidas de desempenho, as letras minúsculas representam probabilidades, enquanto as letras maiúsculas indicam esperanças matemáticas. R indica a taxa de saída de clientes do sistema, que é igual à taxa de chegadas efectiva (isto é, dos clientes a que não é recusada entrada no sistema por limite de capacidade), dado que para um sistema esteja no estado estacionário se verifica que o número médio de entradas é igual ao número médio de saídas.

Desde já se alerta que o símbolo λ pode ter duas interpretações distintas: no caso de modelos com população infinita representa a taxa a que chegam clientes ao sistema, enquanto no caso de população finita representa a taxa de chegada *por cliente*.

Esta notação é a utilizada para quantificar o sistema no estado estacionário. Caso se tratasse de um estado transitório então o tempo t decorrido desde o início do funcionamento do sistema seria um factor importante, e ter-se-ia $p_i(t)$ em vez de p_i, etc.

3.2. Medidas de desempenho no estado estacionário

Um problema que se coloca com frequência é a escolha das medidas de desempenho apresentadas em 3.1 mais relevantes para dada situação. Não há medidas universalmente mais importantes, pois os objectivos de desempenho a atingir variam de aplicação para aplicação e nem sempre são pré-especificados.

Geralmente procura-se uma medida de desempenho que espelhe os custos associados à formação de filas de espera. Por exemplo, o tempo médio que um cliente espera na fila, W_q, ou no sistema, W, tem frequentemente reflexos negativos cujos custos se podem medir.

Noutras aplicações, como por exemplo numa indústria em que os clientes constituam um inventário "aguardando" ser processado ou vendido, o número médio de clientes na fila de espera L_q (tamanho do inventário) pode tornar-se um parâmetro importante. No caso de os clientes serem pessoas que se desloquem ao servidor, a psicologia humana requer que se mantenham os parâmetros W, W_q, e L_q dentro de limites toleráveis.

Outra medida de desempenho muito utilizada é a probabilidade p_i de estarem i clientes no sistema. Por exemplo, em sistemas com capacidade K, a medida p_K indica a probabilidade de a capacidade estar esgotada. No caso particular de sistemas com capacidade K = s, denominados sistemas sem espera ou sistemas de perda, $p_K = b_s$. Nestes sistemas se fala em W_q ou L_q.

A taxa de saída de clientes já servidos, R, indica a quantidade de clientes que sai do sistema por unidade de tempo. Em algumas aplicações essa medida é importante, indicando a produção do sistema. Por exemplo, numa aplicação recente [Allen et al., 1993] foi usado um modelo de fila de espera para prever a quantidade de carvão que determinadas configurações de um sistema de caminho de ferro transportariam de um armazém para uma doca. Este modelo simplificado permitia obter R em função das taxas λ e μ que correspondiam a compromissos entre o número de comboios e o número de carruagens em cada comboio.

Por vezes há aplicações em que se pretende especificar uma medida do nível de serviço oferecido. Interessante sob este aspecto é uma aplicação [Kolesar, 1984], já referida, a servidores ATM localizados numa sala de uma agência bancária. O problema consistia em determinar uma medida de desempenho para várias salas ATM operadas por uma rede bancária, de modo a determinar quais beneficiariam mais de melhoramentos (p. ex. maior número de servidores, servidores mais rápidos, ou alargamento da sala). Foi utilizado o modelo M/M/s/K, dado que os clientes não entravam na sala quando esta estava cheia. Foram calculadas várias medidas de desempenho, tendo sido observado que o tempo médio de espera dos clientes era pequeno mesmo quando a taxa de chegadas era muito grande. Por outro lado, observou-se que muitos clientes se perdiam por estar esgotada a capacidade do sistema. Por esse motivo, passou-se a adoptar p_K como medida da qualidade de serviço em detrimento de W.

3.3. Relações básicas

Existem muitas relações entre as medidas de desempenho no estado estacionário. Algumas medidas representam esperanças matemáticas, podendo ser calculadas pelas relações seguintes:

$$L = \sum_{i=0}^{\infty} i\, p_i \ , \ L_q = \sum_{i=0}^{\infty} i\, q_i = \sum_{i=s}^{\infty} (i-s)\, p_i \ \text{ e } \ B = \sum_{i=0}^{s-1} i\, p_i + \sum_{i=s}^{\infty} s\, p_i$$

Outras relações simples válidas para todos os modelos são:

$L = L_q + B$ (porque B é o número médio de clientes a ser atendido)

$W = W_q + (1/\mu)$ (porque $1/\mu$ é o tempo médio de serviço a um cliente)

$\rho = B / s$ (porque pode definir-se ρ deste modo)

Encontram-se relações válidas apenas para alguns modelos. Por exemplo, $B = 1 - p_0$ para modelos com um único servidor, dado que estarem zero clientes no sistema equivale a ter o servidor desocupado. Nos sistemas que com população e capacidade infinitas verifica-se $R = \lambda$ (porque nenhum cliente é recusado), e $B = \lambda / \mu$, que implica $\rho = \lambda / s\mu$. Nestes sistemas deve verificar-se $\rho < 1$ para que a fila de espera não cresça infinitamente. Nos sistemas com capacidade finita K verifica-se $R = \lambda (1 - p_K)$, dado que os clientes só entram no sistema com probabilidade $(1 - p_K)$. Nestes sistemas a fila de espera não pode crescer infinitamente.

Uma das relações mais conhecidas é a chamada **lei de Little**, que relaciona o número esperado de clientes no sistema com o tempo médio que os clientes estão no sistema através de uma constante de proporcionalidade, a taxa de chegadas efectivas ao sistema. Apesar de não ser imediata como as relações anteriores, apresenta-se sem demonstração (ver demonstração, p. ex., em [Little, 1961]):

$$L = R\, W$$

A lei de Little é extremamente útil porque é genérica. Em particular pode aplicar-se apenas à fila de espera em si, resultando $L_q = R\, W_q$. Estas relações, em conjunto com algumas das acima apresentadas, permitem calcular três das quatro medidas L, W, L_q e W_q, desde que uma destas tenha sido determinada (a medida que seja de mais simples obtenção). No caso particular de sistemas com capacidade infinita a lei de Little pode reescrever-se na sua forma mais habitual:

$$L = \lambda\, W.$$

Habitualmente começa-se por determinar a lei da variável aleatória discreta que indica o número de clientes no sistema no estado estacionário, ou seja, determinar as probabilidades p_i (i=0,1,2,...). A maior parte das medidas de desempenho comuns podem ser calculadas a partir desta distribuição e das relações aqui apresentadas.

4. Modelos de Markov

Habitualmente, só as taxas de chegada e de serviço são conhecidas ou passíveis de ser recolhidas experimentalmente. Por exemplo, muitas vezes deseja-se avaliar o desempenho de sistemas antes de os colocar em funcionamento. Nesses casos há duas vias que se podem seguir: o cálculo dos parâmetros através de soluções analíticas ou a estimação dos parâmetros através de simulação.

A teoria das filas de espera procura encontrar soluções analíticas para os mais variados modelos. Para alguns sistemas é relativamente simples encontrar essas soluções, enquanto para outros ainda não se encontrou alternativa à simulação.

Nesta secção analisa-se uma família de modelos para os quais existe solução analítica: os modelos de Markov. A sua principal característica é a de tanto o processo de chegada como o processo de serviço serem processos de Poisson[6], como por exemplo nos modelos M/M/1 ou M/M/s. Por outras palavras, o tempo entre chegadas e a duração do serviço devem ser variáveis aleatórias reais (v.a.r.) que seguem uma lei exponencial negativa (que será doravante designada por exponencial). Apresentam-se de seguida as principais características desta distribuição, de modo a que se possa verificar quais as situações em que os modelos de Markov melhor se adequam à realidade.

4.1. A distribuição exponencial

A função densidade de probabilidade (f.d.p.) de uma variável aleatória contínua T que segue uma lei exponencial de parâmetro α é (figura A2):

$$f_T(t) = \begin{cases} \alpha\, e^{-\alpha t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

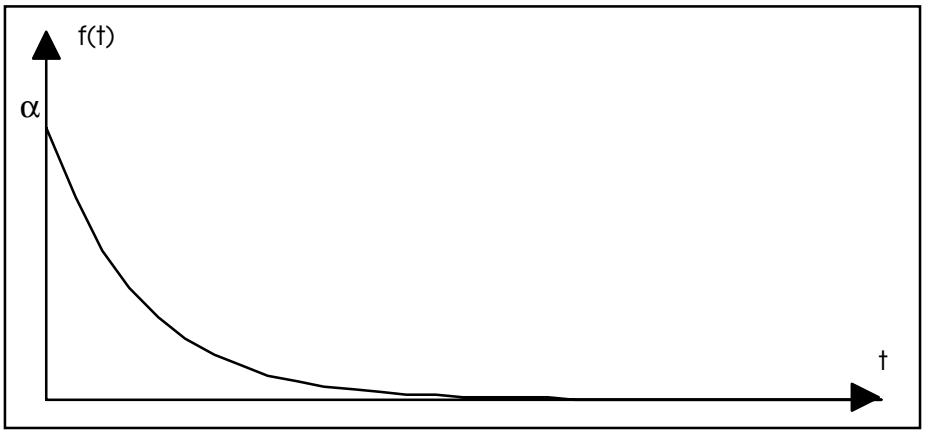


Figura A2 - f. d. p. da lei exponencial

[6] Num processo de Poisson, a sua evolução só depende do estado presente, independentemente do modo como o estado foi adquirido.

A esperança de T é $1/\alpha$, e a sua variância é $1/\alpha^2$. As propriedades seguintes são importantes:

1) f(t) é uma função decrescente em sentido estrito;
2) ausência de memória, ou seja, $P(T > t + x \mid T > x) = P(T > t)$;
3) sejam n v.a.r. independentes $T_1 \sim e(\alpha_1)$, $T_2 \sim e(\alpha_2)$, ... , $T_n \sim e(\alpha_n)$; então $U = \min \{ T_1, T_2, ... , T_n \} \sim e(\alpha)$, com $\alpha = \_ \alpha_i$;
4) as leis exponencial e de Poisson de parâmetro α estão relacionadas, referindo-se a primeira ao tempo entre dois acontecimentos consecutivos e a segunda ao número de acontecimentos por unidade de tempo.

Sempre que se disponha de um histograma, elaborado a partir de dados colhidos por amostragem dos tempos entre chegadas (ou tempos de serviço), deve-se verificar até que ponto este está de acordo com a figura A2 e respeita a propriedade 1. A moda da distribuição é zero, pelo que a chamada classe modal deverá ser a primeira.

No que respeita a distribuição dos tempos entre chegadas, a adopção da lei exponencial é justificável quando os clientes chegam isoladamente, independentemente uns dos outros, e de acordo com um padrão estacionário (isto é, cujos parâmetros não variam no tempo). Muitas vezes é necessário considerar isoladamente períodos distintos para que a condição de estacionariedade seja satisfeita. A condição de independência implica que a chegada de um cliente não é afectada pelas chegadas dos outros, o que exclui por exemplo a possibilidade de um cliente adiar a sua chegada devido à chegada de outro cliente. No caso de clientes chegarem em grupo, a adopção da lei exponencial não se justifica em geral. Pela propriedade 3 é possível utilizar uma lei exponencial para modelar a chegada de clientes de tipos diferentes ignorando as distinções, desde que cada tipo de cliente chegue de acordo com leis exponenciais.

No que concerne a distribuição dos tempos de serviço a justificação da adopção da lei exponencial pode ser mais difícil. De acordo com a figura A2 pode observar-se que se os tempos de serviço forem exponencialmente distribuídos, então serão maioritariamente de curta duração e ocasionalmente de muito longa duração. Por esse motivo, a lei exponencial não se aplica a tempos de serviço cuja duração não varie muito de um valor médio (p. ex. uma tarefa de rotina). Deve considerar-se ainda a razoabilidade da propriedade 2, "ausência de memória", implicada pela adopção da lei exponencial. Essa propriedade indica que não se pode inferir qual a duração remanescente de um serviço já iniciado, a partir da informação da duração já decorrida.

Pode parecer restritivo o universo de aplicação dos modelos de Markov, dadas as condições impostas. Aliás, só muito dificilmente se encontrará uma situação que a lei exponencial possa descrever rigorosamente. Pretende-se apenas alertar para as situações em que a tentação de usar um modelo de Markov deve ser evitada. Contudo, verifica-se que a aproximação é bastante razoável em variadas situações, o que tem conduzido a resultados úteis em muitos casos práticos.

4.2. Soluções analíticas para modelos de Markov simples

Os modelos (de fila de espera) de Markov podem ser resolvidos analiticamente assumindo que os clientes chegam e partem de acordo com um processo de "nascimento e morte"[7]. O termo "nascimento" refere-se a uma chegada de um cliente, enquanto o termo "morte" designa a partida de um cliente após ter sido servido. Esta terminologia deriva da aplicação destes modelos a populações biológicas. Pressupõe-se que só pode ocorrer um nascimento ou morte (uma transição) de cada vez.

O processo pode representar-se por um grafo designado por diagrama de transições ou diagrama de taxas. Na figura A3, por exemplo, cada nó do grafo representa um dos estados de uma fila de espera (cada estado é identificado pelo número de clientes no sistema), enquanto que os arcos indicam as transições possíveis entre estados. Os valores associados a cada arco mostram a taxa média a que as transições ocorrem, assumindo que o sistema atingiu um estado estacionário.

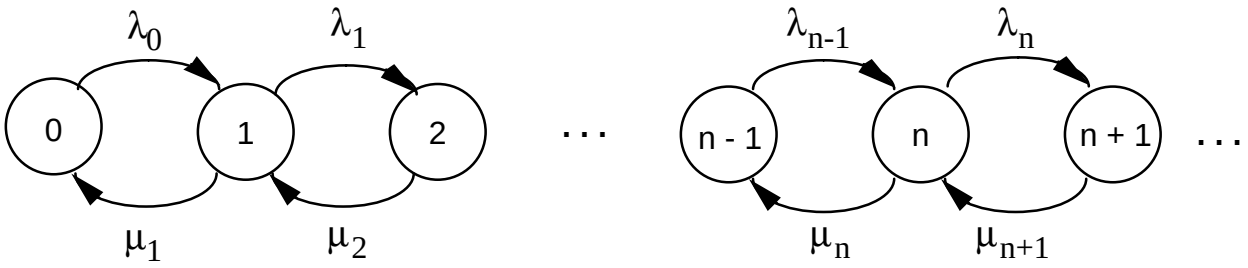


Figura A3 - Diagrama de transições de um processo nascimento-morte

A obtenção de uma solução analítica para as medidas de desempenho no estado estacionário passa pela resolução de um sistema de equações "de balanço"[8]. O princípio subjacente a cada equação de balanço é o de que, no estado estacionário, a taxa média com que um sistema transita para determinado estado tem de ser igual à taxa média com que o sistema transita desse estado para qualquer outro. O sistema de equações lineares a resolver é[9]:

$$\mu_1 p_1 = \lambda_0 p_0$$
$$\lambda_0 p_0 + \mu_2 p_2 = (\mu_1 + \lambda_1) p_1$$
$$\lambda_1 p_1 + \mu_3 p_3 = (\mu_2 + \lambda_2) p_2$$
$$...$$
$$\lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1} = (\mu_n + \lambda_n) p_n$$
$$...$$

A resolução deste sistema de equações é simples. Todas as probabilidades p_i podem ser expressas em função de p_0:

---

7 Estes processos são um caso particular de cadeias de Markov.

8 Também se podem encontrar as designações "equações de equilíbrio estatístico" ou "equações de conservação do fluxo".

9 Estas relações são obtidas a partir das equações de Kolmogorov para cadeias de Markov, quando o sistema se encontra no estado estacionário.

$p_i = C_i\, p_0$, para i = 1, 2, ...

com $C_i = \dfrac{\lambda_{i-1}\, \lambda_{i-2} \ldots \lambda_0}{\mu_i\ \mu_{i-1} \ldots \mu_1}$ para i = 1, 2, ...

Esta expressão é simples de obter a partir do diagrama de transições: para cada estado i, C_i é igual ao produto dos valores dos arcos (superiores) que vão de 0 a i, dividido pelo produto dos valores dos arcos que ligam i a 0.

Como o número de incógnitas é superior em um ao número de equações é necessária uma equação adicional para determinar a solução do sistema. Essa equação é a que deriva do facto de o sistema estar sempre num dos estados possíveis, e de esses estados serem mutuamente exclusivos. A equação é:

$\sum_{i=0}^{\infty} p_i = 1$, que conduz ao resultado $p_0 = \dfrac{1}{1 + \sum_{i=1}^{\infty} C_i}$

Após a determinação dos p_i, o cálculo das diversas medidas de desempenho do sistema no estado estacionário resulta da aplicação das relações já apresentadas na secção 3.3. Apresentam-se de seguida alguns exemplos de obtenção de soluções analíticas para modelos de fila de espera de Markov muito utilizados.

Modelo M/M/1

A figura A4 apresenta o diagrama de transições para um modelo de Markov com um servidor, capacidade infinita e população infinita. A análise do modelo é simplificada pelo facto de as taxas médias de chegada λ e de serviço μ serem constantes.

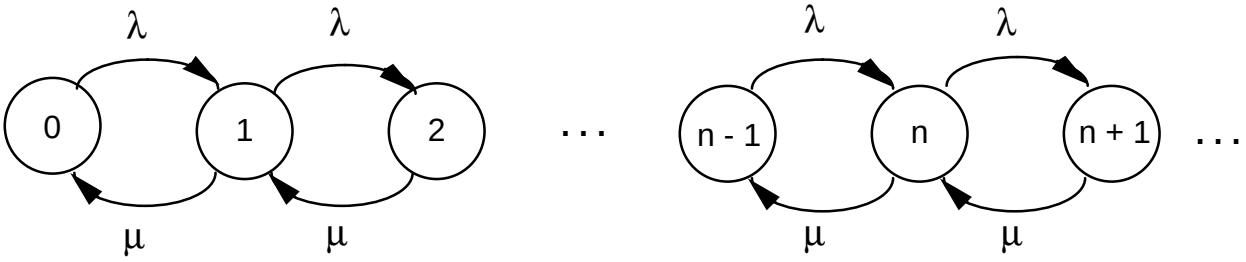


Figura A4 - Diagrama de transições de um modelo M/M/1

A partir da solução do sistema de equações de balanço e das relações existentes no estado estacionário obtém-se:

$C_i = (\lambda / \mu)^i = \rho^i$, para i = 1, 2, ... (dado que, para um sistema uniservidor, $\rho = \lambda / \mu$)

$p_i = \rho^i\, p_0$, para i = 1, 2, ...

$$p_0 = \frac{1}{1+\sum_{i=1}^{\infty}\rho^i} = \frac{1}{\sum_{i=0}^{\infty}\rho^i} = \left(\frac{1}{1-\rho}\right)^{-1} = 1-\rho$$

o que permite escrever $p_i = (1 - \rho)\,\rho^i$, para i = 1, 2, ...

$$L = \sum_{i=0}^{\infty} i(1-\rho)\rho^i = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

$$L_q = \sum_{i=1}^{\infty}(i-1)p_i = \sum_{i=1}^{\infty} ip_i - \sum_{i=1}^{\infty} p_i = L - (1-p_0) = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

A fórmula de Little permitiria então obter W e W_q. O uso todos estes resultados pressupõe que $\mu > \lambda$, caso contrário a fila de espera cresceria sem limite ("explodiria") e o sistema nunca atingiria o estado estacionário. Um resultado talvez inesperado, que resulta de $p_i = (1 - \rho)\,\rho^i$, é o de que, para sistemas M/M/1 capazes de servir todas as chegadas ($\rho<1$), o estado mais provável do sistema é aquele em que o servidor está desocupado e não há clientes na fila!

Modelo M/M/s

No caso de haver vários servidores iguais, o diagrama de transição seria o da figura A5. Sendo μ a taxa média de serviço de cada servidor, as taxas de transição entre os diversos estados já depende do estado do sistema no caso de haver servidores livres. O facto de se ter uma taxa de transição de $n\mu$ quando há n servidores ocupados (n = 1, 2, ..., s) deriva da propriedade 3 da lei exponencial.

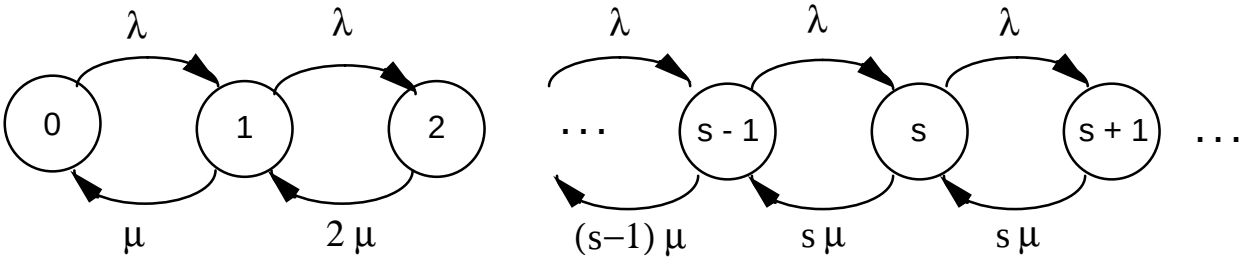


Figura A5 - Diagrama de transições de um modelo M/M/s

A condição para a fila de espera não "explodir" (crescer sem limite) é agora $s\mu > \lambda$ ou, equivalentemente, $\rho < 1$, dado que $\rho = \lambda / s\mu$.

As expressões que se obtêm para as probabilidades p_i são agora mais complexas:

$$C_i = \begin{cases} \dfrac{(\lambda/\mu)^i}{i!}, & \text{para i = 1, 2, ..., s} \\ \dfrac{(\lambda/\mu)^i}{s!\,s^{i-s}}, & \text{para i = s + 1, s + 2, ...} \end{cases} \quad \text{(obtém-se a partir do diagrama de transições)}$$

$$p_i = \begin{cases} \dfrac{(\lambda/\mu)^i}{i!}\, p_0, & \text{para i = 1, 2, ..., s} \\ \dfrac{(\lambda/\mu)^i}{s!\; s^{i-s}}\, p_0, & \text{para i = s + 1, s + 2, ...} \end{cases}$$

$$p_0 = \frac{1}{\sum_{i=0}^{s-1} \dfrac{(\lambda/\mu)^i}{i!} + \dfrac{(\lambda/\mu)^s}{s!} \sum_{i=s}^{\infty} \left(\dfrac{\lambda}{s\mu}\right)^{i-s}} = \frac{1}{\sum_{i=0}^{s-1} \dfrac{(\lambda/\mu)^i}{i!} + \dfrac{(\lambda/\mu)^s}{s!} \dfrac{1}{1-(\lambda/s\mu)}}$$

As expressões para L, L_q, W e W_q podem obter-se pela equação de Little, desde que uma destas seja previamente conhecida. Neste caso a medida mais simples de calcular é L_q, a partir da qual se obtêm as restantes.

$L_q = \sum_{i=s}^{\infty} (i-s) p_i$, expressão que conduz a $\dfrac{p_0 (\lambda/\mu)^s \rho}{s!\,(1-\rho)^2}$

Modelo M/M/s/K (capacidade finita)

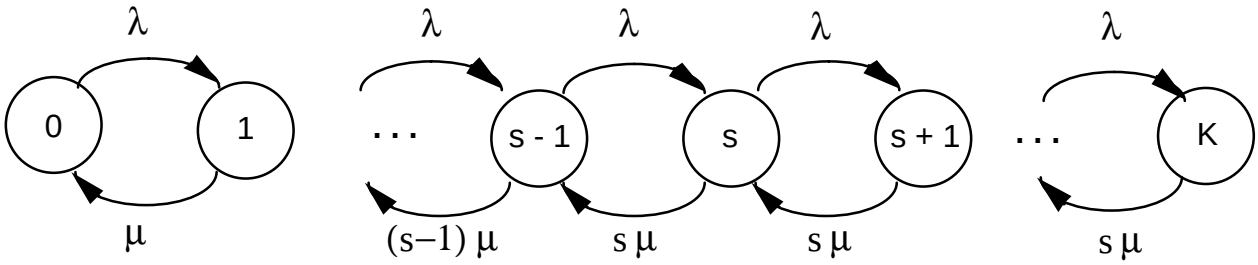


Figura A6 - Diagrama de transições de um modelo M/M/s/K

A figura A6 contém o diagrama de transições para o modelo de Markov com capacidade finita. O número de estados é finito, recusando-se a entrada de novos clientes no sistema sempre que já haja K (K³s) clientes presentes. Pressupõe-se que os clientes cuja entrada é recusada não influenciam o processo de chegada.

O quociente λ / sμ já pode ser superior a 1, uma vez que a fila de espera não pode crescer ilimitadamente. Recorda-se ainda que a fórmula de Little na sua forma habitual L=λW não pode ser usada, devendo ser substituída por L=RW.

A partir do diagrama de transições obtém-se:

$$C_i = \begin{cases} \dfrac{(\lambda/\mu)^i}{i!}, & \text{para i = 1, 2, ..., s} \\ \dfrac{(\lambda/\mu)^i}{s!\; s^{i-s}}, & \text{para i = s + 1, s + 2, ..., K} \end{cases}$$

Outras medidas podem calcular-se a partir destes factores. Omitem-se por serem complexas e poderem ser encontradas em textos sobre este assunto (p. ex. [Hillier e Lieberman, 1990] ou [Jensen, 1986]). Alguns resultados úteis são:
probabilidade de um cliente que chega entrar no sistema: $1 - p_K$
taxa de entrada no (saída do) sistema: $R = \lambda (1 - p_K)$.

As medidas de desempenho obtidas por um modelo M/M/s/K são mais satisfatórias do que as obtidas por um modelo M/M/s, mas à custa de se perder uma fracção p_K do número de potenciais clientes. Como é natural, à medida que K cresce para _ os resultados dos dois modelos aproximam-se.

Modelo M/M/s/N/N (população finita)

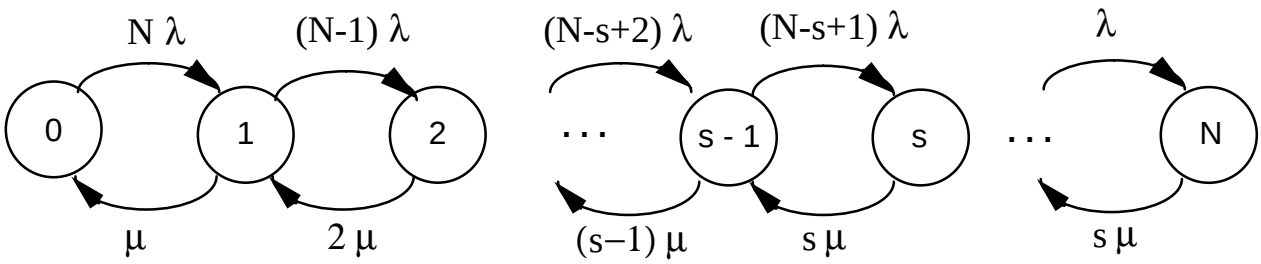


Figura A7 - Diagrama de transições de um modelo M/M/s/N/N

Num sistema com população finita identificam-se N (N³s) potenciais clientes do sistema. Estipula-se uma capacidade N na notação de Kendall, dado que se a capacidade for superior (ou infinita) só será aproveitada N. Cada cliente está em cada momento dentro ou fora do sistema. Quando o cliente está fora do sistema regressa a este após um período que segue uma distribuição exponencial de média 1/λ. Destarte, ao contrário da definição anterior, que se referia à totalidade da população, λ representa a taxa de chegada por cada cliente que está fora do sistema. Quando há n clientes fora do sistema, a taxa de chegadas global é nλ. Este modelo é o tipicamente utilizado para situações em que uma ou várias equipas de reparação (s servidores) é responsável por um conjunto de máquinas (N clientes) que exijam manutenção frequente.

A figura A7 ilustra o diagrama de transições. Deste diagrama retira-se

$$C_i = \begin{cases} \dfrac{N!}{(N-i)!\ i!}\left(\dfrac{\lambda}{\mu}\right)^{\mathrm{i}}, & \text{para i = 1, 2, ..., s} \\ \dfrac{N!}{(N-i)!\ s!\ s^{i-s}}\left(\dfrac{\lambda}{\mu}\right)^{\mathrm{i}}, & \text{para i = s + 1, s + 2, ..., N} \end{cases}$$

A taxa de entrada no (saída do) sistema é $R = \sum_{i=0}^{N} (N-i)\lambda\, p_i$, que é a medida que deve ser usada na fórmula de Little. As restantes medidas seriam obtidas pelo processo habitual.

5. Situações de decisão

Frequentemente, a situação (eventualmente, o problema) que se coloca a uma organização é o de avaliar um sistema de fila de espera e os modos como pode ser melhorado. Outras vezes, pretende-se comparar sistemas alternativos anteriormente à sua colocação em funcionamento. A teoria das filas de espera permite a quem decide (agente de decisão) prever o comportamento de sistemas alternativos a partir de modelos. As variáveis de decisão mais comuns são o número de servidores, a taxa de serviço dos servidores, o número de sistemas de fila de espera e a localização de cada sistema.

Nos casos mais simples, procura-se apenas uma solução que satisfaça determinadas restrições ou obedeça a um único critério de escolha bem definido, como o do menor custo ou do menor tempo de espera médio (secção 5.2). Noutros casos, através de modelos para os custos incorridos pelo desempenho do sistema, procura-se encontrar uma solução que minimize o custo total (secção 5.1). A decisão pode também ser tomada considerando múltiplos critérios, procurando-se então a alternativa que seja mais satisfatória para o agente de decisão. Obviamente nestes casos o resultado poderá variar consoante a valorização (subjectiva) que cada agente de decisão atribui a cada ponto de vista em consideração, dado que frequentemente os critérios são conflituosos entre si (p. ex. pretender maximizar o nível de serviço oferecido e minimizar os custos de serviço simultaneamente). Este é um campo muito rico sobre o qual este texto não se debruçará. Este capítulo termina com uma crítica à aplicação descuidada da teoria das filas de espera, focando alguns sistemas em que os clientes ou os servidores são pessoas (secção 5.3).

5.1. Modelos de optimização de custos

Nesta secção exemplifica-se o modo como a teoria das filas de espera pode auxiliar a encontrar para um sistema uma solução que minimize os custos. Considera-se que há dois tipos de custos: um é directo e tem a ver com os encargos implicados pela oferta de determinado nível de serviço, enquanto outro pode ser directo ou indirecto e reflecte os custos incorridos pela estada dos clientes na fila de espera. Exemplos de custos directos para a organização provocados pela espera em filas são a deterioração de bens voláteis ou o tempo em que funcionários da organização estão em filas como clientes[10] sem produzir (p. ex. à espera de tirar uma fotocópia ou na fila para a máquina de café). Exemplo de um custo indirecto, bastante mais difícil de

[10] Há duas perspectivas para o custo do tempo de um funcionário. Se forem servidores os custos são os associados ao seu salário e restantes regalias. Se forem clientes os custos da permanência em filas de espera são os resultantes do que não se produz pela inactividade do funcionário.

medir, é o que resulta da má imagem com que ficam clientes que a organização serve, e reflecte-se na diminuição da procura[11].

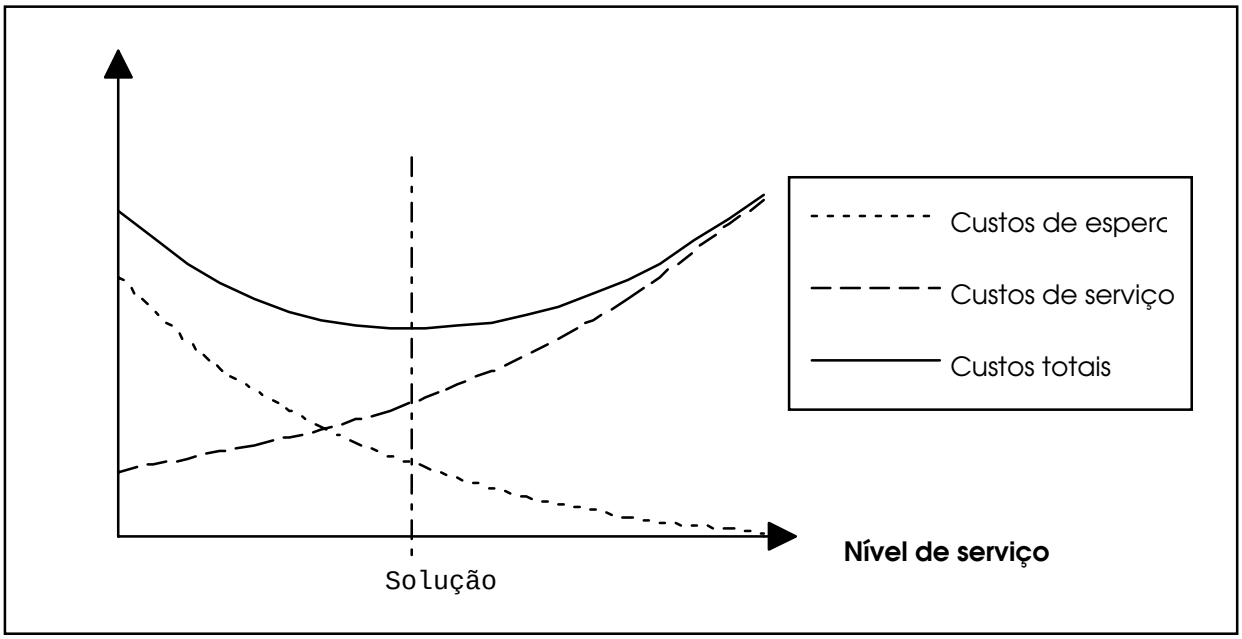


Figura A8 - Obtenção do nível de serviço que minimiza o custo

A obtenção de uma solução que minimize a soma dos dois tipos de custo é esquematizada na figura A8. O custo total esperado $E(C_T)$ é igual ao custo esperado de fornecer um determinado nível de serviço $E(C_S)$, adicionado ao custo esperado pela estada em fila $E(C_E)$ resultante do nível de serviço oferecido:

$$E(C_T) = E(C_S) + E(C_E)$$

Estes custos são tipicamente apresentados como custo por período de tempo. O custo mais difícil de avaliar é $E(C_E)$. Dois modelos simples para o determinar são:

- dependência do número de clientes no sistema, com custo g(i) por período quando estão i clientes no sistema, e p_i como probabilidade de estarem i clientes no sistema

  $$E(C_E) = \sum_{i=0}^{\infty} g(i)\, p_i \,;$$

- dependência do tempo que cada cliente passa no sistema, com custo h(T) por cada cliente que esteja um período T (variável aleatória) no sistema, f(t) como f.d.p. do tempo de espera no sistema T e taxa de chegadas constante λ

  $$E(C_E) = \lambda\, E[h(T)] = \lambda \int_0^{\infty} h(t) f(t) dt$$

[11] Há no entanto organizações que não incorrem neste custo por ausência de competidores. Exemplos típicos são repartições de finanças, conservatórias, tribunais, etc...

Para ambos os modelos resulta $E(C_E) = c_E L$ quando a dependência é linear com um custo c_E por cliente e por período. Como tanto L como $E(C_S)$ podem obter-se a partir dos aspectos que definem o nível de serviço oferecido (nº de servidores, taxa de serviço dos servidores, capacidade do sistema) torna-se possível encontrar a configuração que resulta no mínimo custo total.

Existe, porém, uma variedade de situações em que o custo do tempo de espera não é linear. Um exemplo é o das solicitações para combate a incêndios, dado que os prejuízos causados por incêndios não variam linearmente com o tempo decorrido até à chegada dos bombeiros.

5.2. Escolha de uma solução satisfatória

Muito frequentemente o agente de decisão considera apenas algumas alternativas de melhoramento (ou implantação) de um sistema de fila de espera, escolhendo a alternativa que mais lhe agrade face ao desempenho calculado para os respectivos modelos. Neste caso não se pode dizer que o agente de decisão encontrou a solução óptima[12], uma vez que não analisou todo o conjunto (possivelmente vasto) de possibilidades de melhoramento do seu sistema de fila de espera.

Os resultados da teoria das filas de espera podem ser também utilizados para responder a questões do tipo: "qual a taxa de serviço necessária para manter o tempo de espera na fila abaixo de 5 minutos na horas de maior procura?". Nestes problemas de decisão a taxa de serviço no sistema necessária pode ser obtida utilizando servidores rápidos e/ou utilizando vários servidores. O limiar para o qual se estabelece um limite (superior ou inferior) é tipicamente o tempo de espera no sistema ou na fila, a probabilidade de se perder um cliente (para sistema com capacidade finita), o factor de utilização, ou o número de clientes que está em média no sistema. O limiar é estabelecido através da experiência, de normas existentes, ou por um julgamento subjectivo, não conduzindo necessariamente a um sistema óptimo no que respeita aos custos.

5.3. O factor humano

A escolha de um modelo para determinada situação, bem como a escolha das medidas de desempenho mais adequadas, pode ser uma tarefa complexa quando os clientes são pessoas. Nestas situações emergem factores de ordem psico-sociológica que por um lado originam comportamentos eventualmente difíceis de modelar, e por outro lado dificultam a quantificação do nível de serviço percebido pelos clientes.

---

[12] Diz-se que uma alternativa é óptima no contexto da decisão unicritério se não houver nenhuma outra que a possa superar nesse critério.

Comportamentos já mencionados são o da recusa de entrar no sistema e do abandono da fila de espera por parte dos clientes, para os quais é possível construir modelos. Uma tarefa mais complicada é a de escolher as medidas de desempenho mais importantes para servirem de base à tomada de decisão. Quando os clientes são pessoas, até que ponto uma medida como o tempo de espera médio pode representar a (in)satisfação do cliente com o serviço oferecido?

Um interessante artigo de opinião [Larson, 1987] sugere que a satisfação dos clientes humanos depende de múltiplos atributos e não varia linearmente com o tempo de espera. O próprio tempo de espera é percebido de maneira diferente por clientes diferentes. O aspecto da não-linearidade foi já aflorado no final da secção 5.1 para algumas situações particulares. O carácter multi-atributo da satisfação dos clientes é mais genérico e, de acordo com Larson, depende de três atributos predominantes: justiça do sistema, ambiente da fila de espera e realimentação.

O primeiro atributo, justiça social, pode ser quantificado por uma função que indique a distância entre a ordem pela qual os clientes são servidos e a ordem de chegada. Muitos clientes preferem dirigir-se a um sistema cuja disciplina de serviço seja FCFS em detrimento de outro sistema, ainda que mais rápido (em termos de tempo de espera médio). O segundo atributo, o do ambiente que rodeia a fila de espera, está relacionado com as condições que se oferecem aos clientes enquanto esperam, e tem impacto na percepção da passagem do tempo por parte destes. Por realimentação, o terceiro atributo, entende-se o proporcionar aos clientes estimativas realistas do tempo de espera. Uma outra discussão pertinente em sistemas com clientes (ou servidores) humanos é a da combinação de filas de espera [Rothkopf e Rech, 1987]. Por combinação entende-se a imposição de uma fila única (em vez de uma fila por servidor) em organizações multi-servidor. As soluções analíticas demonstram ser vantajoso combinar sistemas de fila de espera uniservidor independentes num sistema multi-servidor (quando os servidores são todos iguais). Contudo, há de novo fenómenos psico-sociológicos que indicam que em alguns casos pode não ser vantajoso combinar as filas de espera.

Entre os comportamentos que invalidam muitos estudos de combinação de filas estão a escolha a fila mais curta por parte do cliente, e o abandono de uma fila para ingressar numa outra mais curta (ou vazia)[13]. Nestes casos os filas de espera uniservidor deixam de ser independentes, o que dificulta a modelação e a obtenção de resultados. Mesmo que esses comportamentos não ocorram, a manutenção de uma fila de espera por servidor pode ser favorável para o cliente médio no caso de alguns servidores se destinarem a clientes com tempos de serviço mais curtos (p. ex. caixas "expresso" nos hipermercados para clientes com poucos produtos). Um outro factor psicológico contra a formação de uma fila única é o desta ser mais longa (ainda que mais rápida), o que impressiona negativamente os clientes. Contudo, deve-se considerar o argumento de que a formação de uma fila única torna possível servir os clientes por ordem de chegada.

[13] Os tempos de espera médios até podem diminuir no caso de os clientes cujo tempo de serviço seja mais curto poderem mudar de fila mais rapidamente (p. ex. num supermercado são geralmente os clientes com menos compras que mais rapidamente o podem fazer).

No caso em que os servidores também sejam pessoas os argumentos para não combinar filas de espera são em maior número. O mais importante será talvez o facto de, em certas organizações, os servidores trabalharem mais rapidamente quando se sentem responsáveis (e pressionados) pela "sua" fila de espera, do que se forem apenas co-responsáveis por uma fila única. Outros argumentos para não combinar filas são a oferta ao cliente da possibilidade de escolher o servidor (fortalecendo desse modo os laços entre o cliente e o servidor), e a possibilidade de atribuir aos servidores tarefas especializadas (de modo a permitir um serviço mais rápido).

6. Outros modelos de fila de espera

Os modelos de Markov constituem uma ferramenta de grande utilidade e uma razoável aproximação à realidade para muitas situações. No entanto, a existência de elementos contrariando os pressupostos dos modelos de Markov conduz algumas vezes à adopção de modelos mais complexos. As possibilidades de modelação são inúmeras, mas só algumas permitem a obtenção de resultados analíticos. Apresentam-se a título ilustrativo alguns dos resultados mais conhecidos.

6.1. Disciplina de serviço

Os resultados obtidos até agora assumiam a disciplina em que os clientes são atendidos por ordem de chegada (FCFS). Serão estes válidos caso a disciplina de serviço seja outra?

A resposta é sim, para os casos em que a selecção do próximo cliente a ser servido não seja influenciada por uma estimativa do seu tempo de serviço. De facto, os diagramas de transições foram elaborados sem atender a qualquer disciplina de serviço em particular. Dado que os resultados foram obtidos unicamente a partir de tais diagramas, estes permanecem válidos para disciplinas como a de servir primeiro o último cliente (LCFS) ou a de seleccionar os clientes ao acaso (RSS).

Apesar das medidas de desempenho mais comuns permanecerem iguais, a disciplina de serviço pode afectar as distribuições de variáveis aleatórias como por exemplo a do tempo de espera dos clientes. Os clientes esperam na fila esperam **em média** o mesmo tempo W_q, mas a variância pode ser diferente. Mostra-se que a disciplina FCFS é a que proporciona variâncias menores, enquanto a LCFS resulta na maior variância e a RSS produzirá um valor intermédio. Os esquemas com prioridades permitem reduzir os tempos de espera para alguns clientes seleccionados, à custa dos restantes.

Um caso inesperado onde se experimentou utilizar as disciplinas RSS e LCFS é o do atendimento de emergências médicas ou policiais críticas. A causa desta estratégia é o facto do atendimento ser muito mais eficaz quando o tempo de espera é pequeno. Ainda que se cometam

"injustiças", a probabilidade de se prenderem criminosos em flagrante delito é maior ao acorrer primeiro às chamadas mais recentes do que se seguir a ordem de chamada.

Quando existem estimativas para o tempo de serviço de cada cliente (p. ex. em aplicações industriais) pode encontrar-se uma disciplina pontual que procure optimizar determinado parâmetro de desempenho. O estudo desses métodos, ditos de escalonamento[14] ou sequenciação, está, contudo, fora do âmbito deste texto.

6.2. Modelos não markovianos

A análise dos modelos não markovianos é consideravelmente dificultada. Porém, encontram-se frequentemente casos em que os processos de chegada ou, sobretudo, de serviço não são processos de Poisson. Felizmente, tem sido possível obter resultados de grande utilidade.

Existem soluções analíticas, por exemplo, para os sistemas $M/E_k/1$ e $E_k/M/1$, dado que uma v.a.r. que segue uma lei Erlang-k é equivalente à soma de k v.a.r. exponencialmente distribuídas. É possível por esse motivo construir-se um diagrama de transições que, não sendo do tipo de um processo nascimento-morte, é uma cadeia de Markov para a qual se podem obter resultados. Outro estratagema para obter resultados é o de encontrar cadeias de Markov discretas "embebidas" nos sistemas, ao retratar o sistema em instantes seleccionados, ignorando o comportamento do sistema entre esses instantes. Isso é possível para o modelo G/M/1, considerando apenas os instantes em que os clientes chegam, e para o modelo M/G/1 considerando os momentos em que os serviços terminam. Para o modelo M/G/1 tal procedimento conduz a um resultado conhecido pela fórmula de Pollaczek-Khinchine:

$$L = \frac{\rho^2 + \lambda^2 V(S)}{2(1-\rho)} + \rho = \frac{2\rho - \rho^2 + \lambda^2 V(S)}{2(1-\rho)}$$, o que implica

$$L_q = L - \rho = \frac{\rho^2 + \lambda^2 V(S)}{2(1-\rho)}$$ pelas relações apresentadas em 3.3.

Nestas relações $\rho=\lambda/\mu$, e V(S) indica a variância da distribuição S dos tempos de serviço. Não é necessário conhecer completamente essa distribuição, mas apenas a sua média $(1/\mu)$ e a sua variância. Este é um resultado muito conveniente que se aplica a qualquer modelo com processo de chegada de Poisson e capacidade e população infinitas. Em particular, para o modelo M/D/1, em que os tempos de serviço são determinísticos, V(S)=0, pelo que

[14] O termo método escalonamento designa, neste contexo, um algoritmo ou uma heurística para impor uma ordem de serviço a um conjunto de clientes que espera ser atendido. O mesmo termo é também utilizado para designar algoritmos ou heurísticas que visam determinar o modo como alguns recursos se repartem para satisfazer determinada procura, tendo em conta eventuais restrições (p. ex. elaboração dos horários escolares).

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

Uma interessante comparação pode ser efectuada com o modelo M/M/1, em que S ~ e(μ), e $V(S)=1/\mu^2$, o que conduz ao resultado já conhecido

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

Pode observar-se, comparando estas duas fórmulas, que o tempo de espera médio é reduzido para metade quando os tempos de serviço deixam de ser aleatórios "puros" (exponencialmente distribuídos) e passam a ser constantes (determinísticos). De algum modo, pode atribuir-se metade do tempo de espera na fila à variância nos tempos de serviço, atribuindo-se o restante à variância nos tempos de chegada[15]. Qualquer modelo $M/E_K/1$ tem medidas de desempenho entre as obtidas por M/D/1 (limite "superior") e M/M/1 (limite "inferior").

Da fórmula de Pollaczek-Khinchine também se pode concluir que um servidor pode compensar uma menor taxa de serviço média E(S)=μ por menor variância V(S), na comparação com um servidor mais rápido (maior μ), mas também mais inconstante (maior variância).

Um outro resultado interessante é o das fórmulas para o cálculo de p_0, p_1, ..., p_s para o modelo M/M/s/s, de fácil obtenção, serem generalizáveis ao modelo M/G/S/S sem qualquer alteração.

6.3. Redes de filas de espera

Em várias aplicações, por exemplo na área industrial ou de transportes, podem existir vários sistemas de fila de espera em rede, saindo os clientes de um sistema para entrar noutro ou para deixar a rede (figura A9). Estabelecem-se assim redes de filas de espera. Nestas redes, os clientes passam por uma sequência de sistemas de fila de espera para receberem serviços distintos. Essa sequência de visitas pode ser ou não idêntica para todos os clientes: cada cliente pode seguir o seu próprio percurso.

[15] Num sistema D/D/1 em que ρ<1 nunca se formaria uma fila de espera.

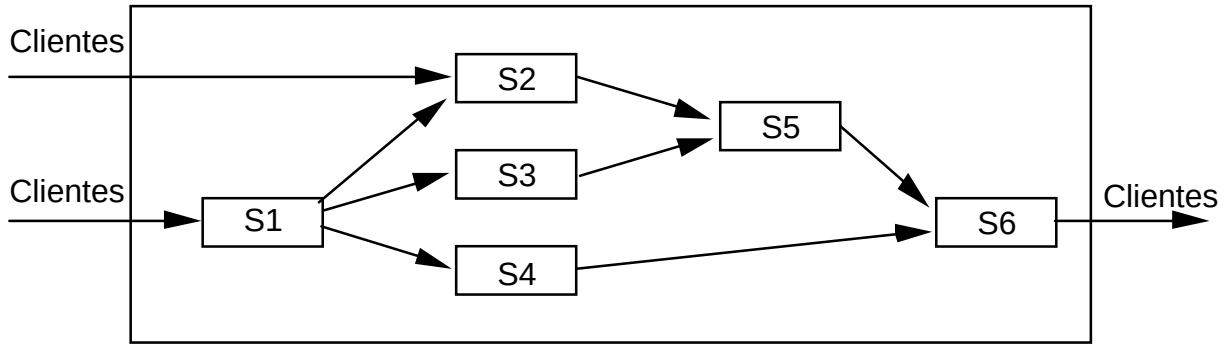


Figura A9 - Exemplo de uma rede de filas de espera

A análise destas organizações pode ser bastante complexa dada a interacção existente entre os diversos sistemas de fila de espera. Existe, porém, uma classe de redes de filas de espera cuja análise é simplificada, desde que satisfaçam algumas condições:

1. Quaisquer fluxos de clientes que cheguem do exterior da rede de filas de espera ocorrem de acordo com processos de Poisson independentes.
2. Os tempos de serviço são exponencialmente distribuídos, e as respectivas taxas de serviço dependem apenas da fila de espera local a cada sistema.
3. A capacidade de cada sistema de fila de espera é infinita.
4. Para cada sistema i, os clientes transitam para um sistema j com probabilidade p_{ij} ou abandonam a rede com probabilidade q_i.

A análise das redes de filas de espera que satisfazem as condições anteriores[16] é possível devido à seguinte propriedade:

propriedade de equivalência - Num sistema M/M/s, cujo processo de chegada é de Poisson com taxa λ e o processo de serviço para cada servidor é de Poisson com taxa μ, e tal que $s\mu>\lambda$, a saída dos clientes ocorre segundo um processo de Poisson com taxa λ.

As redes satisfazendo as condições 1 a 4 têm uma solução, que se obtém analisando cada sistema de fila de espera pertencente à rede isoladamente. Devido às condições 1 e 2, e também devido às propriedades da distribuição exponencial, cada sistema da fila de espera pode ser analisado pelo modelo M/M/s.

6.4. Obtenção de resultados

Quando se constrói um modelo de fila de espera é necessário considerar o modo como se obtêm resultados. Já se mostrou que para a importante classe dos sistemas de Markov é simples obter soluções analíticas. Na secção 6.1 deram-se exemplos de alguns outros sistemas para os quais se conhecem soluções analíticas. Nesta secção apresentam-se alguns exemplos de outras vias para a obtenção das medidas de desempenho desejadas.

16 São por vezes designadas por redes de Jackson.

Uma via para obtenção de resultados é ainda analítica. Considere-se por exemplo o modelo G/G/1[17], para o qual não há solução analítica exacta. Existem, no entanto, expressões que proporcionam aproximações aceitáveis, ou que indicam limites inferiores e superiores para algumas medidas de desempenho. No caso de modelos com vários servidores, por exemplo M/G/s, é por vezes possível encontrar modelos inferiores ou superiores com um só servidor, para os quais exista solução analítica (p. ex. o modelo M/G/1). Para um sistema com n servidores todos iguais podem obter-se resultados "optimistas" através de um modelo com um único servidor cuja taxa de serviço seja n vezes superior.

A outra via é a da simulação, que visa estimar estatisticamente os parâmetros pretendidos. A simulação constitui uma matéria importante no contexto das filas de espera[18]. A sua força reside no facto de se poderem modelar através da simulação sistemas muito complexos, cuja análise seria difícil ou mesmo impossível de outra forma. A sua fraqueza consiste no facto de os resultados obtidos não fornecem soluções exactas, mas estimativas, que variam frequentemente de simulação para simulação.

Habitualmente a simulação efectua-se utilizando computadores. Os chamados simuladores são programas capazes de efectuar simulações, como por exemplo o GPSS, sendo apenas necessário parametrizar a situação. Em alternativa, existem linguagens de simulação, como o SIMSCRIPT ou o MODSIM, que permitem elaborar complexos programas de simulação mais específicos para os fins em vista. Os custos incorridos pelo recurso à simulação são geralmente superiores aos de uma abordagem analítica.

Por vezes a simulação complementa o resultado da utilização de outros modelos. Numa aplicação recente [Allen et al., 1993], tentou-se melhorar o funcionamento de um sistema de transporte de carvão de um armazém para um porto marítimo. A ligação era ferroviária, devendo cada comboio esperar para ser carregado no armazém, seguir para o porto para descarregar, e regressar para novo carregamento. A motivação para o estudo efectuado foi a lentidão com que os navios no porto eram carregados. As alternativas a considerar eram cerca de 900, correspondendo a diversas combinações de parâmetros como o número de carruagens por comboio e o número de comboios. Apesar de o sistema ser demasiado complexo para ser modelado fielmente por uma fila de espera, e se tenha optado por abordagens de simulação e de escalonamento, utilizou-se um modelo de fila de espera para efectuar um primeiro rastreio de alternativas promissoras.

O servidor foi considerado o armazém, sendo o tempo de serviço igual ao tempo de carregamento do comboio. Usou-se um modelo M/M/1/N/N para cada alternativa, variando N (número de comboios), a taxa de chegada e a taxa de serviço (ambas crescentes com N), tendo sido obtidos resultados muito úteis e corroborados pela simulação posteriormente efectuada.

---

[17] Esta notação surge frequentemente como GI/G/1, em que GI indica tempos de chegada arbitrariamente distribuídos e independentes.

[18] Nos livros de investigação operacional é habitual encontrar-se um capítulo sobre simulação a par de um capítulo sobre a teoria das filas de espera.

Como curiosidade, refira-se que os resultados apontaram para um aumento no número de comboios e diminuição dos seus tamanhos, tendo o desempenho do sistema melhorado consideravelmente após a adopção das recomendações do estudo.

REFERÊNCIAS E BIBLIOGRAFIA CONSULTADA

[Allen et al., 1993] Allen G., E. Gunn e P. Rutherford, Improving throughput of a coal transport system with the aid of three simple models, *Interfaces*, Vol. 23, Jul-Ago 1993, pp. 88-103.

[Anderson et. al., 1991] Anderson, D. R., D. J. Sweeney e T. A. Williams, *An introduction to management science*, 6ª ed., West Publishing Co., St. Paul.

[Cooper, 1981] Cooper, Robert B., *Introduction to queuing theory*, North Holland.

[Hillier e Lieberman, 1990] Hillier, Frederick S. e Gerald J. Lieberman, *Introduction to operations research*, 5ª ed., McGraw-Hill, New York.

[Jensen, 1986] Jensen, Paul A., *Students guide to operations research*, McGraw-Hill, New York.

[Kolesar, 1984] Kolesar, Peter, Stalking the endangered CAT: a queueing analysis of congestion at automatic teller machines, *Interfaces*, Vol. 14, Nov-Dez 1984, pp. 16-26.

[Lane et al., 1993] Lane, M. S., A. H. Mansour e J. L. Harpell, Operations research techniques: a longitudinal update 1973-1988, *Interfaces*, Vol. 23, Mar-Abr 1993, pp. 63-68.

[Larson, 1987] Larson, R. C., Perspectives on queues: social justice and the psychology of queueing, *Operations Research*, Vol. 35, 1987, pp. 895-905.

[Little, 1961] Little, J. D. C., A proof for the queueing formula: $L = \lambda W$, *Operations Research*, Vol. 9, pp. 383-387.

[Ravindran, 1987] Ravindran, A., D. Phillips, J. Solberg, *Operations research - principles and practice*, 2ª ed., John Wiley & Sons, Chichester.

[Rothkopf e Rech, 1987] Rothkopf, M. H. e P. Rech, Perspectives on queues: combining queues is not always beneficial, *Operations Research*, Vol. 35, 1987, pp. 906-908.

[Turban e Meredith, 1981] Turban, E. e J. R. Meredith, *Fundamentals of management science*, Business Publications Inc., Plano.

OUTRA BIBLIOGRAFIA RELEVANTE PARA O TEMA

[Gross e Harris, 1974] Gross, D. e C. M. Harris, *Fundamentals of queueing theory*, Wiley, New York.

[Kleinrock, 1975] Kleinrock, L., *Queueing systems, Vol. I*, Wiley, New York.

[Kleinrock, 1976] Kleinrock, L., *Queueing systems, Vol. II*, Wiley, New York.

BIBLIOGRAFIA PARA A DISCIPINA

[Hillier e Lieberman, 1990] Hillier, Frederick S. e Gerald J. Lieberman, *Introduction to operations research*, 5ª ed., McGraw-Hill, New York.

[Ravindran, 1987] Ravindran, A., D. Phillips, J. Solberg, *Operations research - principles and practice*, 2ª ed., John Wiley & Sons, Chichester.