# INTRODUÇÃO À PESQUISA OPERACIONAL

**8ª Edição**

**FREDERICK S. HILLIER**
*Stanford University*

**GERALD J. LIEBERMAN**
*Ex-Professor Titular da Stanford University*

**Tradução**
ARIOVALDO GRIESI

**Revisão Técnica**
JOÃO CHANG JUNIOR
Doutor em Administração — FEA/USP
Professor Titular do Programa de Mestrado da UNIP
Professor Titular da FAAP

Bangcoc Bogotá Beijing Caracas Cidade do México
Cingapura Lisboa Londres Madri Milão Montreal Nova Delhi
Santiago São Paulo Seul Sydney Taipé Toronto

CAPÍTULO 17

# Teoria das Filas

As *filas* (filas de espera) fazem parte do dia-a-dia de nossa vida. Todos nós esperamos em uma fila para: comprar o ingresso para uma sessão de cinema, fazer um depósito bancário, pagar as compras em um supermercado, remeter um pacote no correio, comprar um sanduíche em uma lanchonete, brincar em um parque de diversões etc. Acabamos nos acostumando a um volume considerável de espera, mas ainda assim nos irritamos se tivermos de aguardar muito em uma fila.

Entretanto, ter de esperar não se limita apenas a esses transtornos pessoais de relativa insignificância. O tempo que a população de um país perde em filas é um importante fator tanto na qualidade de vida nesse país quanto na eficiência da economia dessa nação. Por exemplo, antes de sua dissolução, a União Soviética era notória por filas enormes que seus cidadãos freqüentemente tinham de suportar para comprar suas necessidades básicas. Mesmo nos Estados Unidos, estima-se que os norte-americanos gastem 37.000.000.000 horas por ano esperando em filas. Se, no entanto, esse tempo fosse gasto produtivamente, resultaria em aproximadamente 20 milhões de pessoas-ano de trabalho útil!

Mesmo esse número absurdo não é capaz de representar todo o impacto de se causar uma espera excessiva. Grandes ineficiências também ocorrem por causa de outros tipos de espera, além daquelas de pessoas esperando em uma fila. Por exemplo, deixar *máquinas* esperando para serem reparadas pode resultar em perdas na produção. *Veículos* (inclusive navios e caminhões) que precisam aguardar para ser descarregados podem atrasar embarques seguintes. *Aviões* aguardando para decolar ou pousar podem afetar horários de vôos posteriores. Atrasos em transmissões de *telecomunicações* devido a linhas saturadas podem provocar problemas técnicos com os dados. Fazer que *ordens de produção* fiquem esperando para ser realizadas pode afetar a produção de lotes seguintes. Realizar *serviços* após a data combinada pode resultar na perda de futuros negócios.

A *teoria das filas* é o estudo da espera em todas essas formas diversas. Ela usa *modelos de filas* para representar os diversos tipos de *sistemas de filas* (sistemas que envolvem filas do mesmo tipo) que surgem na prática. As fórmulas para cada modelo indicam como o sistema de filas correspondente deve funcionar, inclusive o tempo de espera médio que ocorrerá, em uma série de circunstâncias.

Portanto, esses modelos de filas são muito úteis para determinar como operar um sistema de filas da forma mais eficiente. Fornecer capacidade de atendimento em excesso para operar o sistema envolve custos demasiados. Porém, não fornecer capacidade de atendimento suficiente resulta em espera excessiva e todas suas lamentáveis conseqüências. Os modelos permitem encontrar um equilíbrio apropriado entre custo de serviço e o tempo de espera.

Após uma discussão geral sobre o assunto, o presente capítulo apresenta a maioria dos modelos de filas elementares e seus resultados básicos. A Seção 17.10 discute como as informações fornecidas pela teoria das filas pode ser usada para elaborar sistemas de filas que minimizem o custo total de serviço e de espera e, a seguir, o Capítulo 26 (no CD-ROM) fornece mais detalhes sobre a aplicação da teoria das filas dessa maneira.

## 17.1 EXEMPLO-PROTÓTIPO

A sala de emergências do HOSPITAL MUNICIPAL atende a casos de emergência, fornecendo os devidos cuidados médicos, que chegam ao hospital em ambulâncias ou em carros particulares. A qualquer hora existe um médico de plantão na sala de emergências. Entretanto, em virtude de uma tendência crescente de esses casos de "emergência" para usar essas instalações em vez de irem ao consultório médico particular, o hospital tem passado por um aumento contínuo no número de atendimentos na sala de emergências a cada ano. Conseqüentemente, tornou-se bastante comum pacientes chegarem durante horas de pico (no início da noite) e terem de esperar até chegar a sua vez de ser atendido pelo médico. Por essa razão, foi feita uma proposta de se alocar um segundo médico para a sala de emergências durante esse horário de pico, de modo que duas emergências pudessem ser atendidas ao mesmo tempo. O administrador do hospital foi designado para estudar essa questão.[1]

O administrador começou a coletar dados históricos relevantes e depois projetou-os para o ano seguinte. Reconhecendo que a sala de emergências é um sistema de filas, ele aplicou diversos modelos alternativos da teoria das filas para prever as características de espera do sistema com um e dois médicos, como pode ser observado nas seções posteriores deste capítulo (ver Tabelas 17.2 e 17.3).

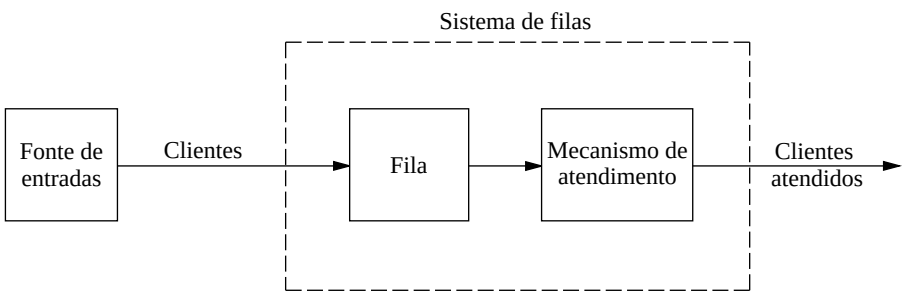
## 17.2 ESTRUTURA BÁSICA DOS MODELOS DE FILAS

### Processo de Filas Básico

O processo básico suposto pela maioria dos modelos de filas é o seguinte. *Clientes* que necessitam de atendimento chegam ao longo do tempo por uma *fonte de entradas.* Esses clientes entram no *sistema de filas* e pegam uma *fila.* Em certos momentos, um membro da fila é selecionado para atendimento por alguma regra conhecida como *disciplina da fila.* O atendimento necessário é então realizado para o cliente pelo *mecanismo de atendimento,* após o qual o cliente deixa o sistema de filas. Esse processo é representado na Figura 17.1.

Podem ser feitas diversas hipóteses alternativas sobre os vários elementos do processo de filas; elas serão discutidas a seguir.

FIGURA 17.1
Processo de filas básico.




[1] Para um estudo de caso real desse tipo, ver BOLLING, W. Blaker. Queueing Model of a Hospital Emergency Room. *Industrial Engineering,* p. 26-31, set. 1972.

### Fonte de Entradas (População Solicitante)

Uma característica da fonte de entradas é o seu tamanho. O *tamanho* é o número total de clientes que poderiam precisar de atendimento de tempos em tempos, isto é, o número total de possíveis clientes distintos. Essa população de onde provêm as chegadas é conhecida como **população solicitante**. O tamanho pode ser suposto como *infinito* ou *finito* (de modo que a fonte de entradas também seja dita *ilimitada* ou *limitada*). Como os cálculos são bem mais fáceis para o caso infinito, normalmente parte-se dessa hipótese mesmo quando o tamanho real for um número finito relativamente grande; e ela deve ser assumida como hipótese implícita para qualquer modelo de filas que não afirme o contrário. O caso finito é mais difícil analiticamente, pois o número de clientes no sistema de filas afeta o número de possíveis clientes fora do sistema a qualquer momento. Entretanto, deve ser feita a hipótese finita caso a taxa na qual a fonte de entradas gere clientes novos seja significativamente afetada pelo número de clientes no sistema de filas.

O padrão estatístico pelos quais os clientes chegam ao longo do tempo também deve ser especificado. A hipótese comum é que eles chegam de acordo com um *processo de Poisson*; isto é, o número de clientes que chegam até dado momento tem uma distribuição de Poisson. Conforme discutido na Seção 17.4, esse caso é aquele na qual as chegadas ao sistema de filas ocorrem aleatoriamente, porém, a certa taxa média fixa, independentemente de quantos clientes já se encontrarem lá (de forma que o *tamanho* da fonte de entradas seja *infinito*). Uma hipótese equivalente é que a distribuição probabilística do tempo entre as chegadas consecutivas é uma distribuição *exponencial*. As propriedades dessa distribuição são descritas na Seção 17.4. O tempo entre as chegadas consecutivas é conhecido como **tempo entre chegadas**.

Quaisquer hipóteses incomuns sobre o comportamento de clientes que chegam também devem ser especificadas. Um exemplo é a *recusa*, na qual o cliente se recusa a entrar no sistema e será perdido caso a fila seja muito longa.

### Fila

A fila é o local onde os clientes aguardam *antes* de ser atendidos. Uma fila é caracterizada pelo número máximo de clientes permitidos que ela pode conter. As filas são chamadas *infinitas* ou *finitas*, conforme esse número for infinito ou finito. A hipótese de uma *fila infinita* é o padrão para a maioria dos modelos de filas, mesmo para situações em que ele realmente é um limite superior finito (relativamente grande) sobre o número de clientes permitido, pois lidar com um limite superior destes seria um fator complicador na análise. Entretanto, para sistemas de filas em que esse limite superior for suficientemente pequeno, de modo que ele seria efetivamente atingido com alguma freqüência, torna-se necessário supor uma *fila finita.*

### Disciplina da Fila

A disciplina da fila se refere à ordem na qual membros da fila são selecionados para atendimento. Ela poderia ser, por exemplo, os primeiros que chegam serão os primeiros a ser atendidos, aleatória, de acordo com algum procedimento de prioridade ou algum outro tipo de ordem. Normalmente, para modelos de filas adota-se o critério dos primeiros que chegam serão os primeiros a ser atendidos, a menos que seja combinado de outra forma.

### Mecanismo de Atendimento

O mecanismo de atendimento é formado por uma ou mais *instalações de atendimento*, cada uma das quais contendo um ou mais *canais de atendimento paralelos*, chamados **atendentes**. Se existir mais de uma instalação de atendimento, o cliente poderá ser atendido por uma seqüência destes (*canais de atendimento em série*). Em dada instalação, o cliente entra em um desses canais de atendimento paralelos e é completamente atendido por esse atendente. Um modelo de filas deve especificar a disposição das instalações e o número de atendentes (canais paralelos) em cada uma delas. A maioria dos modelos elementares parte

do pressuposto de uma instalação de atendimento com um atendente ou com um número finito de atendentes.

O tempo decorrido entre o início do atendimento até o seu término para um cliente em uma instalação de atendimento é denominado **tempo de atendimento** (ou *tempo de permanência*). Um modelo de determinado sistema de filas deve especificar a distribuição probabilística de tempos de atendimento para cada atendente (e, possivelmente, para tipos diferentes de clientes), embora seja comum supor a *mesma* distribuição para todos os atendentes (todos os modelos neste capítulo partem desse pressuposto). A distribuição de tempo de atendimento que é suposta com maior freqüência na prática (em grande parte porque ela é bem mais fácil de ser tratada) é a distribuição *exponencial* discutida na Seção 17.4 e a maioria de nossos modelos é desse tipo. Outras distribuições de tempo de atendimento importantes são a distribuição *degenerada* (tempo de atendimento constante) e a distribuição de *Erlang* (gama), conforme ilustrado pelos modelos na Seção 17.7.

### Processo de Filas Elementar

Conforme já sugerido, a teoria das filas foi aplicada a vários tipos diferentes de situações com filas de espera. Entretanto, o tipo mais freqüente de situação é o seguinte: uma fila de espera única (que, às vezes, pode estar vazia) se forma na frente de uma única instalação de atendimento, dentro da qual se encontram um ou mais atendentes. Cada cliente que chega pela fonte de entradas é atendido por um dos atendentes, talvez após algum tempo aguardando na fila (fila de espera). O sistema de filas envolvido é representado na Figura 17.2.

Note que o processo de filas no exemplo ilustrativo da Seção 17.1 é desse tipo. A fonte de entradas gera clientes na forma de casos de emergência necessitando de cuidados médicos. Uma sala de emergências é a instalação de atendimento e os médicos são os atendentes.

Um atendente não precisa ser um único indivíduo, ele pode ser um grupo de pessoas, por exemplo, uma equipe de manutenção que combina forças para realizar simultaneamente o serviço exigido para um cliente. Além disso, os atendentes não precisam nem mesmo ser pessoas. Em muitos casos, um atendente pode ser, em vez disso, uma máquina, um veículo, um dispositivo eletrônico etc. Da mesma maneira, os clientes na fila de espera não precisam, necessariamente, ser pessoas. Eles poderiam, por exemplo, ser peças aguardando por certa operação a ser executada por determinado tipo de máquina ou então carros aguardando em frente de uma cabine de pedágio.

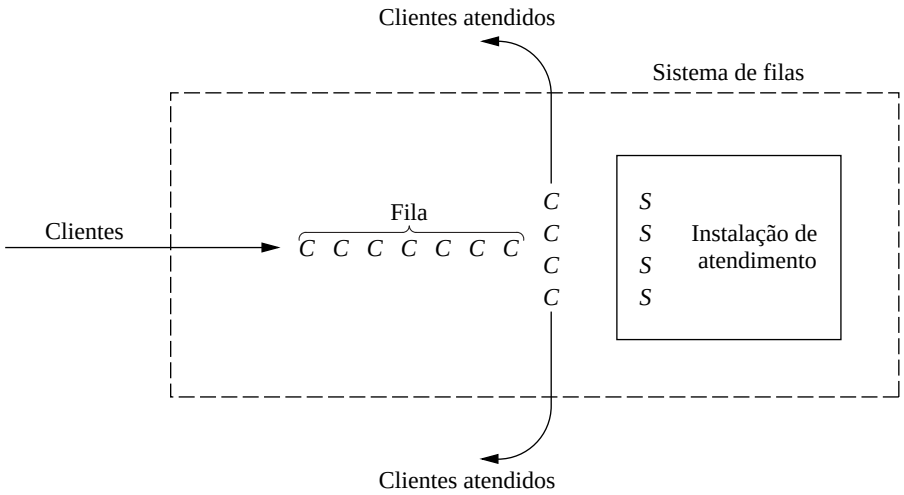
Não é necessário que haja, na verdade, uma fila de espera física formada em frente de uma estrutura física que componha a instalação de atendimento. Os membros da fila poderiam estar espalhados por certa área, esperando que um atendente chegue até eles, por exemplo, máquinas aguardando para serem consertadas. O atendente ou grupo de atendentes alocados para determinada área forma a instalação de atendimento para aquela área. A teoria das filas ainda fornece o número médio de esperas, o tempo de espera médio e assim por diante, pois é irrelevante se os clientes esperam juntos em um grupo. A única exigência essencial para a teoria das filas ser aplicável é que mudanças no número de clientes aguardando por dado serviço ocorrem da mesma forma que a situação física descrita na Figura 17.2 (ou um equivalente legítimo) predomina.

Exceto pela Seção 17.9, todos os modelos de filas discutidos neste capítulo são do tipo elementar representado na Figura 17.2. Muitos desses modelos supõem, além disso, que todos os *tempos entre chegadas* sejam independentes e distribuídos de forma idêntica e que todos os *tempos de atendimento* sejam independentes e distribuídos de forma idêntica. Tais modelos são identificados convencionalmente como se segue:

Distribuição de tempos de atendimento

– / – / – Número de atendentes

Distribuição de tempos entre atendimentos,




■ FIGURA 17.2
Um sistema de filas elementar (cada cliente é indicado por um *C* e cada atendente por um *S*).

em que $M$ = distribuição exponencial (markoviana), conforme descrito na Seção 17.4,
$D$ = distribuição degenerada (tempos constantes), conforme discutido na Seção 17.7,
$E_k$ = distribuição de Erlang (parâmetro de forma = $k$), conforme descrito na Seção 17.7,
$G$ = distribuição geral (qualquer distribuição arbitraria permitida),[2] conforme discutido na Seção 17.7.

Por exemplo, o modelo *M*/*M*/*s* discutido na Seção 17.6 parte do pressuposto de que tanto os tempos entre atendimentos quanto os tempos de atendimento possuem uma distribuição exponencial e que o número de atendentes é *s* (qualquer inteiro positivo). O *modelo M/G/1* discutido novamente na Seção 17.7 parte do pressuposto de que os tempos entre atendimentos possuem uma distribuição exponencial, porém ele não coloca nenhuma restrição sobre qual deve ser a distribuição de tempos de atendimento, ao passo que o número de atendentes restringe-se exatamente a 1. Na Seção 17.7, também são introduzidos vários outros modelos que caem nesse esquema de identificação.

### Terminologia e Notação

A menos que declarado de outra forma, será adotado o seguinte padrão em termos de terminologia e notação:

Estado do sistema = número de clientes no sistema de filas.
Comprimento da fila = número de clientes aguardando que o atendimento se inicie
= estado do sistema *menos* número de clientes que estão sendo atendidos.
$N(t)$ = número de clientes no sistema de filas no instante $t$ ($t \geq 0$).
$P_n(t)$ = probabilidade de exatamente $n$ clientes se encontrarem no sistema de filas no instante $t$, dado o número no instante 0.
$s$ = número de atendentes (canais de atendimento paralelos) no sistema de filas.

[2] Quando nos referimos a tempos entre atendimentos, é convenção substituir o símbolo *G* por *GI* = distribuição independente geral.

$\lambda_n$ = taxa média de chegada (número de chegadas esperado por unidade de tempo) de novos clientes quando $n$ clientes se encontram no sistema.

$\mu_n$ = taxa média de atendimento para o sistema global (número de clientes esperado completando o atendimento por unidade de tempo) quando $n$ clientes se encontram no sistema. *Nota:* $\mu_n$ representa a taxa *combinada* na qual todos os atendentes *ocupados* (aqueles que se encontram atendendo clientes) completam o atendimento.

$\lambda$, $\mu$, $\rho$ = ver o parágrafo a seguir.

Quando $\lambda_n$ for uma constante para todo $n$, essa constante é representada por $\lambda$. Quando a taxa média de atendimento *por atendente ocupado* for uma constante para todo $n \geq 1$, essa constante é representada por $\mu$. Nesse caso, $\mu_n = s\mu$ quando $n \geq s$, isto é, quando todos os $s$ atendentes estiverem ocupados. Sob essas condições, $1/\lambda$ e $1/\mu$ são, respectivamente, o *tempo esperado entre atendimentos* e o *tempo de atendimento esperado*. Da mesma forma, $\rho = \lambda/(s\mu)$ é o **fator de utilização** para a instalação de atendimento, isto é, a fração de tempo esperada em que atendentes individuais se encontram ocupados, pois $\lambda/(s\mu)$ representa a fração da capacidade de atendimento ($s\mu$) do sistema que está sendo *utilizada* em média pelos clientes que chegam ($\lambda$).

Também é necessária certa notação para descrever resultados *de estado estável*. Quando um sistema de filas começar a operar recentemente, o estado do sistema (número de clientes no sistema) será afetado enormemente pelo estado inicial e pelo tempo que passou. Diz-se que o sistema se encontra em uma **condição transitória**. Entretanto, após ter decorrido um tempo suficiente, o estado do sistema se torna basicamente independente do estado inicial e o tempo decorrido (exceto sob circunstâncias incomuns).[3] O sistema basicamente atingiu agora uma **condição de estado estável**, na qual a distribuição probabilística do estado do sistema permanece a mesma (a distribuição de *estado estável* ou *estacionária*) ao longo do tempo. A teoria das filas teve uma tendência de se concentrar em grande parte na condição de estado estável, em parte porque o caso transitória é mais difícil analiticamente. Existem alguns resultados transitória, mas eles geralmente vão além do escopo técnico deste livro. A notação indicada a seguir supõe que o sistema se encontre em uma *condição de estado estável:*

$P_n$ = probabilidade de exatamente $n$ clientes se encontrarem no sistema de filas.

$L$ = número de clientes esperado no sistema de filas $= \sum_{n=0}^{\infty} nP_n$.

$L_q$ = comprimento esperado da fila (exclui clientes que estão sendo atendidos)
$= \sum_{n=s}^{\infty} (n - s)P_n$.

$\mathcal{W}$ = tempo de espera no sistema (inclui o tempo de atendimento) para cada cliente individual.

$W = E(\mathcal{W})$.

$\mathcal{W}_q$ = tempo de espera na fila (exclui o tempo de atendimento) para cada cliente individual.

$W_q = E(\mathcal{W}_q)$.

### Relações entre *L*, *W*, *Lq* e *Wq*

Suponha que $\lambda_n$ seja uma constante $\lambda$ para todo $n$. Foi provado que em um processo de filas de estado estável,

[3] Quando $\lambda$ e $\mu$ são definidos, essas circunstâncias incomuns são que $\rho \geq 1$, em cujo caso o estado do sistema tende a ficar cada vez maior à medida que o tempo passa.

$$L = \lambda W.$$

Pelo fato de John D. C. Little[4] ter obtido a primeira prova rigorosa, essa equação algumas vezes é chamada **fórmula de Little**. Além disso, a mesma prova também demonstra que

$$L_q = \lambda W_q.$$

Se $\lambda_n$ não forem iguais, então $\lambda$ pode ser substituído nessas equações por $\overline{\lambda}$ a taxa *média* de chegada a longo prazo. Iremos mostrar posteriormente como $\overline{\lambda}$ pode ser determinado para alguns casos básicos.

Suponha agora que o tempo médio de atendimento seja constante, $1/\mu$ para todo $n \geq 1$. Segue então que

$$W = W_q + \frac{1}{\mu}.$$

Essas relações são extremamente importantes, pois elas permitem que possam ser determinadas imediatamente todas as quatro quantidades fundamentais — $L$, $W$, $L_q$ e $W_q$ — assim que uma delas for encontrada analiticamente. Essa situação é oportuna, por que algumas dessas quantidades normalmente são muito mais fáceis de ser encontradas que outras quando um modelo de filas é solucionado a partir de princípios básicos.

## 17.3 EXEMPLOS DE SISTEMAS DE FILAS REAIS

Nossa descrição de sistemas de filas na Seção 17.2 pode parecer relativamente abstrata e aplicável somente a situações práticas muito especiais. Pelo contrário, os sistemas de filas são surpreendentemente freqüentes em ampla gama de contextos. Para ampliar nossos horizontes sobre a aplicabilidade da teoria das filas, iremos mencionar brevemente vários exemplos de sistemas de filas reais que caem em diversas categorias amplas. A seguir descreveremos sistemas de filas em diversas empresas proeminentes (além de uma prefeitura) e estudos de caso renomados que foram conduzidos para desenvolver esses sistemas.

### Algumas Classes de Sistemas de Filas

Uma importante classe de sistemas de filas que todos nós encontramos em nossas vidas diárias são os **sistemas de atendimento comercial**, em que clientes externos recebem atendimento de organizações comerciais. Muitas delas envolvem atendimento pessoa a pessoa em um local permanente, como em uma barbearia (os barbeiros são os atendentes), caixas em um banco, caixas em uma loja e uma fila de lanchonete (canais de serviço em série). Entretanto, muitas outras não se enquadram nessas condições, como conserto de certos eletrodomésticos (em que o atendente vai até o cliente), uma máquina automática de vendas (o atendente é uma máquina) e um posto de gasolina (os carros são os clientes).

Outra classe importante é a dos **sistemas de atendimento de transporte**. Para alguns desses sistemas, os veículos são os clientes, por exemplo, carros aguardando em um posto de pedágio ou em um semáforo (o atendente), um caminhão ou navio esperando ser carregado ou descarregado por uma equipe (os atendentes) e aviões aguardando para pousar ou decolar de uma pista (o atendente). Um exemplo incomum desse tipo é o de um estacionamento, onde os carros são os clientes e as vagas, os atendentes, porém não há nenhuma fila, pois os clientes que chegam vão para outro lugar para estacionar caso a vaga esteja ocupada. Em outros casos, os veículos, como táxis, caminhões de bombeiros e elevadores são os atendentes.

[4] LITTLE, J. D. C. A Proof for the Queueing Formula: $L = \lambda W$. *Operations Research*, v. 9, n. 3, p. 383-387, 1961; ver também STIDHAM, JR., S. A Last Word on $L = \lambda W$. *Operations Research*, v. 22, n. 2, p. 417-421, 1974.

Nos últimos anos, a teoria das filas provavelmente foi aplicada mais a **sistemas de atendimento interno**, em que os clientes recebendo atendimento são *internos* à organização. Entre os exemplos podemos citar sistemas de manipulação de materiais, nos quais as unidades de manipulação de materiais (os atendentes) deslocam cargas (os clientes); sistemas de manutenção, em que as equipes de manutenção (os atendentes) consertam máquinas (os clientes) e estações de inspeção, onde inspetores de controle de qualidade (os atendentes) inspecionam itens (os clientes). Instalações de funcionários e departamentos atendendo outros funcionários também caem nessa categoria. Além disso, máquinas podem ser vistas como atendentes cujos clientes são as tarefas que estão sendo processadas. Um exemplo relacionado seria o de um laboratório de computadores, onde cada computador é visto como o atendente.

Há um reconhecimento crescente hoje em dia de que a teoria das filas também pode ser aplicada a **sistemas de serviços sociais**. Por exemplo, um sistema judicial é uma rede de filas, em que os tribunais são instalações de atendimento, os juízes (ou painéis de juízes) são os atendentes e os processos aguardando julgamento, os clientes. Um sistema legislativo é uma rede de filas similar, na qual os clientes são os projetos de lei aguardando aprovação. Diversos sistemas de assistência médica também são sistemas de filas. Já vimos um exemplo na Seção 17.1 (uma sala de emergências de um hospital), mas poderíamos também interpretar ambulâncias, aparelhos de raios X e camas de um hospital como atendentes em seus próprios sistemas de filas. Similarmente, famílias esperando por sistemas habitacionais de custo baixo ou moderado ou outros serviços sociais podem ser vistos como clientes em um sistema de filas.

Embora estas sejam quatro classes abrangentes de sistemas de filas, elas não esgotam a lista. De fato, a teoria das filas começou no início deste século com aplicações para a telefonia (o fundador da teoria das filas, A. K. Erlang, foi funcionário da Cia. Telefônica Dinarmaquesa em Copenhagen) e a telefonia ainda é uma aplicação importante. Além disso, todos nós temos nossas filas pessoais — tarefas domésticas, livros a serem lidos e assim por diante. Entretanto, esses exemplos são suficientes para sugerir que sistemas de filas de fato invadem muitas áreas da sociedade.

### Alguns Estudos Renomados para Desenvolver Sistemas de Filas

O prestigioso *Franz Edelman Awards for Management Science Achievement* é uma premiação concedida anualmente pelo Institute of Operations Research and Management Sciences (Informs) para as melhores aplicações de PO do ano. Um número bastante substancial dessas premiações foi concebido a aplicações inovadoras da teoria das filas no desenvolvimento de sistemas de filas. Descrevemos brevemente algumas dessas aplicações a seguir.

Um dos primeiros ganhadores (descrito na edição de novembro de 1975, Parte 2, da *Interfaces*) foi a *Xerox Corporation.* A empresa introduziu recentemente um importante sistema de cópia que estava demonstrando ser de extrema valia para seus usuários. Conseqüentemente, esses clientes estavam exigindo que os técnicos de campo da Xerox reduzissem o tempo de espera para reparar essas máquinas. Uma equipe de PO aplicou então a teoria das filas para estudar como melhor atender as novas exigências de atendimento. Isso resultou na substituição das zonas de atendimento anteriores com um técnico de campo por zonas com três técnicos. Essa mudança teve um impacto drástico tanto na redução substancial dos tempos médios de espera dos clientes quanto no aumento da utilização dos técnicos de campo em 50%.

Na Seção 3.5, descrevemos uma renomada aplicação da *United Airlines* (edição de janeiro de 1986 da *Interfaces*) que resultou em uma economia anual de mais de US$ 6 milhões. Essa aplicação envolvia programar as escalas de 4.000 agentes de reservas e pessoal de suporte da United em seus 11 escritórios de reservas e 1.000 agentes de atendimento a clientes em seus dez maiores aeroportos. Após determinar quantos empregados seriam necessários em cada local durante cada meia hora da semana, discutimos como a programação linear foi aplicada para desenvolver as escalas para todos os empregados visando atender essas exigências de atendimento de forma mais eficiente. Entretanto, jamais mencionamos

como foram estabelecidas essas exigências de atendimento sobre o número de empregados necessários cada meia hora.

Agora, estamos em condições de destacar que essas exigências de atendimento foram determinadas aplicando-se a *teoria das filas.* Cada local específico (por exemplo, os balcões de *check-in* em um aeroporto) forma um sistema de filas com os empregados sendo os atendentes. Após prever a taxa média de chegada durante cada meia hora da semana, foram usados modelos de filas para encontrar o número mínimo de atendentes que forneceriam medidas de desempenho satisfatórias para o sistema de filas.

A *L.L. Bean, Inc.*, a maior empresa de telemarketing e vendas por correio, baseou-se principalmente na teoria das filas para seu renomado estudo de como alocar seus recursos de telecomunicações. As informações do artigo, descrevendo esse estudo, se encontram na edição de janeiro de 1991 da *Interfaces* e outros artigos dando informações adicionais se encontram nas edições de novembro de 1989 e de março-abril de 1993 desse jornal. As chamadas telefônicas provenientes de seu *call center* para fazer pedidos são os clientes em um grande sistema de filas, com os agentes de telemarketing como atendentes. As questões-chave durante o estudo foram as seguintes.

1. Quantas linhas-tronco deveriam ser disponibilizadas para telefonemas que chegam no *call center*?
2. Quantos agentes de telemarketing deveriam ser alocados em vários horários?
3. Quantas linhas de espera deveriam ser fornecidas para clientes aguardando um agente de telemarketing? Note que o número limitado de linhas de espera faz que o sistema tenha uma fila finita.

Para cada interessante combinação dessas três quantidades, modelos de filas fornecem as medidas de desempenho do sistema de filas. Dadas essas medidas, a equipe de PO avaliou cuidadosamente o custo de vendas perdidas em razão de alguns clientes encontrarem linha ocupada ou serem colocados em uma linha de espera por muito tempo. Acrescentando-se o custo de recursos de telemarketing, a equipe foi capaz de encontrar a combinação de três quantidades que minimiza o custo total esperado. Isso resultou em uma economia de custos de cerca de US$ 9 a US$ 10 milhões por ano.

A cidade de *Nova York* tem uma longa e permanente tradição de usar técnicas de PO em planejar e operar muitos de seus complexos sistemas de atendimento urbanos. Iniciando no final dos anos de 1960, estudos renomados, envolvendo a teoria das filas, foram conduzidos pelo seu Corpo de Bombeiros e seu Departamento de Polícia. Emergências policiais e de incêndio são os clientes nesses respectivos sistemas de filas. Subseqüentemente, importantes estudos de PO (incluindo diversos casos mais complexos envolvendo a teoria das filas) foram conduzidos por seu Departamento Sanitário, Departamento de Transportes, Departamento de Saúde Pública, Departamento de Proteção Ambiental, Gabinete de Administração e Orçamento e Departamento de Suspensão Condicional de Penas Judiciais. Em razão do sucesso desses estudos, muitos deles agora têm suas próprias equipes internas de PO.

O renomado estudo da cidade de Nova York que iremos descrever aqui envolve seu sistema de detenção à acusação. Esse sistema é formado pelo processo iniciando na prisão de indivíduos até eles serem acusados (o primeiro comparecimento no tribunal perante um juiz de acusação, que determina se houve ou não uma causa provável para a prisão). Antes desse estudo, os detidos na cidade (os clientes em um sistema de filas) ficavam em custódia aguardando julgamento por uma média de 40 horas (eventualmente mais de 70 horas). Esses tempos de esperas foram considerados excessivos, porque os detidos eram mantidos em ambientes ruidosos e abarrotados de gente que eram emocionalmente estressantes, insalubres e muitas vezes fisicamente perigosos. Portanto, foi realizado um estudo de PO de dois anos para revisar o sistema. Foram empregadas tanto a teoria das filas quanto a simulação (tema do Capítulo 20). Isso levou a mudanças operacionais e de políticas de grande extensão que reduziram simultaneamente o tempo médio de espera até a acusação a 24 horas ou menos e geraram uma economia anual de US$ 9,5 milhões. Ver a edição de janeiro de 1993 da *Interfaces* para mais detalhes.

O primeiro prêmio na edição de 1993 foi ganho pela AT&T por um estudo que (como o precedente) também combinasse o emprego da teoria das filas e simulação (edição janeiro-fevereiro de 1994 da *Interfaces*). Os modelos de filas referem-se tanto à rede de telecomunicações da AT&T quanto para o *call center* para clientes comerciais típicos da AT&T. O propósito do estudo foi o de desenvolver um sistema amigável baseado em PC que os clientes comerciais da AT&T podem usar para orientá-los no desenho ou redesenho de seus *call centers*. Já que os *call centers* formam um dos mercados de maior crescimento nos Estados Unidos, esse sistema foi usado cerca de 2.000 vezes pelos clientes comerciais da AT&T desde 1992. Isso resultou em uma economia superior a US$ 750 milhões em lucros anuais para esses clientes.

A *KeyCorp* é uma das maiores empresas controladoras bancárias nos Estados Unidos, com mais de 1.300 agências e mais de 6.000 caixas. O renomado estudo de PO dessa empresa (edição de janeiro de 1996 da *Interfaces*) concentrou-se no emprego da teoria das filas para aumentar o desempenho do sistema de filas de cada agência onde os caixas atendem os clientes. Isso resultou no desenvolvimento de um sistema de gerenciamento de excelência em serviços (SEMS) para toda a empresa. Uma parte fundamental do SEMS é um sistema de captura de desempenho que coleta dados de forma contínua para cada componente discreto de cada transação no caixa em um processo completamente automatizado. Esse sistema permite ao SEMS medir atividades das agências e gerar relatórios sobre os tempos de espera dos clientes, níveis de produtividade e de competência dos caixas. Esses relatórios ajudam os gerentes a programar a escala de caixas para se adequar melhor às chegadas dos clientes. Eles também identificam oportunidades para melhorar a produtividade e o atendimento fornecido pelos caixas redesenhando o processo de atendimento e fornecendo padrões de desempenho. Esses esforços levaram a uma imensa redução de 53% nos tempos médios de atendimento, uma grande melhoria nos tempos de espera por parte dos clientes e um importante aumento no nível de satisfação do cliente. Ao mesmo tempo, espera-se que o SEMS reduza despesas com pessoal em US$ 98 milhões ao longo de cinco anos.

A *Hewlett-Packard* (HP) é um fabricante multinacional de equipamentos eletrônicos líder de mercado. Em 1993, a empresa instalou um sistema de linha de montagem mecanizado para a fabricação de impressoras jato de tinta em seu complexo fabril em Vancouver, Washington, para atender à explosiva demanda por tal tipo de impressora. Assim, tornou-se aparente que o sistema instalado não seria suficientemente rápido ou confiável para atender às metas de produção da empresa. Portanto, foi constituída uma equipe conjunta de cientistas da administração da HP e do Massachusetts Institute of Technology (MIT) para estudar como redesenhar o sistema para melhorar seu desempenho.

Conforme descrito na edição janeiro-fevereiro de 1998 da *Interfaces* para esse célebre estudo ganhador de premiações, a equipe HP/MIT rapidamente percebeu que o sistema de linha de montagem poderia ser modelado como um tipo especial de sistema de filas no qual os clientes (as impressoras a serem montadas) passariam por uma série de atendentes (operações de montagem) em uma seqüência fixa. Um modelo de filas especial para esse tipo de sistema gerou rapidamente os resultados analíticos que foram necessários para determinar como o sistema deveria ser redesenhado para alcançar a capacidade exigida da forma mais econômica possível. As mudanças incluíam acrescentar maior espaço de armazenagem em pontos estratégicos para manter melhor o fluxo de trabalho a estações subseqüentes e para minimizar o efeito de falhas de máquina. O novo *design* aumentou a produtividade em cerca de 50% e gerou um aumento nas receitas de aproximadamente US$ 280 milhões em vendas de impressoras, bem como receitas adicionais de produtos secundários. Essa aplicação inovadora do modelo de filas especial também deu à HP um novo método para criar projetos de sistemas rápidos e eficientes posteriormente em outras áreas da empresa.

Existem muitas outras aplicações premiadas da teoria das filas para o projeto de sistemas de filas, bem como inúmeros artigos adicionais descrevendo outras aplicações bem-sucedidas. Entretanto, os diversos exemplos apresentados nesta seção felizmente lhe deram uma idéia dos tipos de aplicações que estão ocorrendo e do impacto que algumas vezes eles têm.

## 17.4 O PAPEL DA DISTRIBUIÇÃO EXPONENCIAL

As características operacionais dos sistemas de filas são determinadas, em grande parte, por duas propriedades estatísticas, a saber, a distribuição probabilística dos *tempos entre atendimentos* (ver "fonte de entradas" na Seção 17.2) e a distribuição probabilística dos *tempos de atendimento* (ver "Mecanismos de Atendimento" na Seção 17.2). Para sistemas de filas reais, essas distribuições podem assumir praticamente qualquer forma. A única restrição é que não podem ocorrer valores negativos. Entretanto, para formular um *modelo* de teoria das filas como uma representação do sistema real, é necessário especificar a forma assumida de cada uma dessas distribuições. Para ser útil, a forma assumida deveria ser *suficientemente realista* cujo modelo fornece *previsões razoáveis* enquanto, ao mesmo tempo, ser *suficientemente simples* cujo modelo é *matematicamente tratável.* Baseado nessas considerações, a distribuição probabilística mais importante na teoria das filas é a *distribuição exponencial.*

Suponha que uma variável aleatória $T$ represente tempos entre chegadas ou tempos de atendimento. Iremos nos referir às ocorrências que marcam o final desses tempos — chegadas ou finalizações de atendimentos — como *eventos.* Diz-se que essa variável aleatória tem uma distribuição exponencial *com parâmetro* $\alpha$ se sua função de densidade probabilística for

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

conforme mostrado na Figura 17.3. Nesse caso, as probabilidades acumulativas são

$$\begin{aligned} P\{T \leq t\} &= 1 - e^{-\alpha t} \\ P\{T > t\} &= e^{-\alpha t} \end{aligned} \qquad (t \geq 0),$$

e o valor esperado e a variância de $T$ são, respectivamente,

$$E(T) = \frac{1}{\alpha},$$

$$\text{var}(T) = \frac{1}{\alpha^2}.$$

Quais são as implicações de se supor que $T$ possui uma distribuição exponencial para um modelo de filas? Para explorar essa questão, examinemos seis propriedades fundamentais da distribuição exponencial.

**Propriedade 1:** $f_T(t)$ é uma função estritamente *decrescente* de $t$ ($t \geq 0$).

Uma conseqüência da Propriedade 1 é que

$$P\{0 \leq T \leq \Delta t\} > P\{t \leq T \leq t + \Delta t\}$$

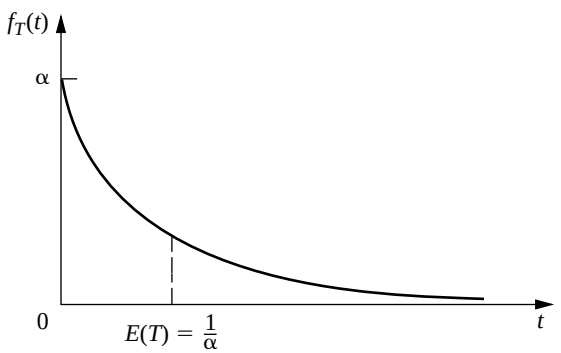



FIGURA 17.3
Função de densidade probabilística para a distribuição exponencial.

para quaisquer valores estritamente positivos de $\Delta t$ e $t$. Essa conseqüência é decorrente do fato de que essas probabilidades são a área abaixo da curva $f_T(t)$ ao longo do intervalo de comprimento $\Delta t$ indicado e a altura média da curva é menor para a segunda probabilidade que para a primeira. Portanto, não é somente possível, mas também relativamente provável, que $T$ assumirá um pequeno valor próximo de zero. De fato,

$$P\left\{0 \le T \le \frac{1}{2}\frac{1}{\alpha}\right\} = 0{,}393$$

ao passo que

$$P\left\{\frac{1}{2}\frac{1}{\alpha} \le T \le \frac{3}{2}\frac{1}{\alpha}\right\} = 0{,}383.$$

de modo que o valor que $T$ assume é mais provável que seja "pequeno" [isto é, menos da metade de $E(T)$] do que "próximo" ao seu valor esperado [isto é, não muito além da metade de $E(T)$], embora o segundo intervalo seja o dobro do primeiro.

Essa é uma propriedade razoável para $T$ em um modelo de filas? Se $T$ representa *tempos de atendimento*, a resposta depende da natureza geral do atendimento envolvido, conforme discutido a seguir.

Se o atendimento necessário for basicamente idêntico para cada cliente, com o atendente sempre realizando a mesma seqüência de operações de atendimento, então os tempos de atendimento reais tendem a estar próximos do tempo de atendimento esperado. Podem ocorrer pequenos desvios em relação à média, mas normalmente em decorrência de pequenas variações na eficiência do atendente. Um pequeno tempo de atendimento muito longe da média é praticamente impossível, pois é preciso certo tempo mínimo para realizar as operações de atendimento necessárias mesmo quando o atendente está trabalhando em alta velocidade. A distribuição exponencial claramente não fornece uma boa aproximação para a distribuição de tempos de atendimento para esse tipo de situação.

No entanto, considere uma situação na qual as tarefas específicas necessárias do atendente diferem entre os diversos tipos de clientes. A natureza abrangente do atendimento pode ser a mesma, porém o tipo específico e o tempo de atendimento diferem. Por exemplo, este seria o caso do problema da sala de emergências do Hospital Municipal discutido na Seção 17.1. Os médicos se deparam com ampla gama de problemas clínicos. Na maioria dos casos, eles podem oferecer o tratamento necessário de forma bem rápida, contudo, eventualmente um paciente pode exigir tratamento mais intensivo. De maneira similar, caixas de bancos e caixas de lojas são outros atendentes desse tipo geral, em que o atendimento necessário normalmente é breve, todavia, eventualmente, pode ser mais demorado. Uma distribuição exponencial de tempos de atendimento seria bastante plausível para esse tipo de situação de atendimento.

Se $T$ representar *tempos entre atendimentos*, a Propriedade 1 descarta situações nas quais possíveis clientes que se aproximem do sistema de filas tendam a adiar sua entrada, caso vejam outro cliente que está à sua frente. Entretanto, é totalmente consistente com o fenômeno comum das chegadas ocorrerem "aleatoriamente", descrito por propriedades subseqüentes. Portanto, quando tempos de chegada forem colocados em uma linha de tempo, eles algumas vezes têm a aparência de estar concentrados com eventuais intervalos grandes separando essas concentrações, em razão da grande probabilidade de tempos entre atendimentos pequenos e a pequena probabilidade de tempos entre atendimentos grandes, mas um padrão irregular como este faz parte da aleatoriedade.

**Propriedade 2:** Falta de memória.

Essa propriedade pode ser declarada matematicamente como

$$P\{T > t + \Delta t \,|\, T > \Delta t\} = P\{T > t\}$$

para quaisquer valores positivos $t$ e $\Delta t$. Em outras palavras, a distribuição probabilística do tempo *remanescente* até o evento (chegada ou término do atendimento) ocorrer é sempre a mesma, independentemente de quanto tempo ($\Delta t$) já tiver passado. De fato, o processo se "esquece" de seu passado. Esse surpreendente fenômeno acontece com a distribuição exponencial, pois

$$\begin{aligned} P\{T > t + \Delta t \,|\, T > \Delta t\} &= \frac{P\{T > \Delta t,\, T > t + \Delta t\}}{P\{T > \Delta t\}} \\ &= \frac{P\{T > t + \Delta t\}}{P\{T > \Delta t\}} \\ &= \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha \Delta t}} \\ &= e^{-\alpha t} \\ &= P\{T > t\}. \end{aligned}$$

Para *tempos entre atendimentos*, essa propriedade descreve a situação corriqueira na qual o tempo até a próxima chegada não sofre nenhuma influência de quando ocorreu a última chegada. Para *tempos de atendimento*, a propriedade é mais difícil de ser interpretada. Não deveríamos esperar que ela fosse respeitada em uma situação em que o atendente tem de realizar a mesma seqüência fixa de operações para cada cliente, pois então um atendimento de longa duração implicaria que provavelmente pouco restaria a ser feito. Entretanto, no tipo de situação na qual as operações de atendimento necessárias diferem entre os clientes, a declaração matemática da propriedade pode ser bastante realista. Para esse caso, se um tempo de atendimento considerável já tivesse decorrido para um cliente, a única implicação poderia ser que esse cliente em particular precisaria de atendimento mais amplo que a maioria.

**Propriedade 3:** O *mínimo* de diversas variáveis aleatórias exponenciais independentes tem uma distribuição exponencial.

Para declarar essa propriedade matematicamente, façamos que $T_1, T_2, \ldots, T_n$ sejam variáveis aleatórias exponenciais *independentes* com parâmetros $\alpha_1, \alpha_2, \ldots, \alpha_n$, respectivamente. Façamos também que $U$ seja a variável aleatória que admita o valor igual ao *mínimo* dos valores realmente assumidos por $T_1, T_2, \ldots, T_n$; isto é,

$$U = \text{mín}\,\{T_1, T_2, \ldots, T_n\}.$$

Portanto, se $T_i$ representar o tempo até que determinado tipo de evento ocorra, então $U$ representará o tempo até que o *primeiro* dos $n$ eventos diversos ocorra. Observe agora que para qualquer $t \geq 0$,

$$\begin{aligned} P\{U > t\} &= P\{T_1 > t, T_2 > t, \ldots, T_n > t\} \\ &= P\{T_1 > t\}P\{T_2 > t\} \cdots P\{T_n > t\} \\ &= e^{-\alpha_1 t} e^{-\alpha_2 t} \cdots e^{\alpha_n t} \\ &= \exp\left(-\sum_{i=1}^{n} \alpha_i t\right), \end{aligned}$$

de modo que $U$ de fato tenha uma distribuição exponencial com parâmetro

$$\alpha = \sum_{i=1}^{n} \alpha_i.$$

Essa propriedade apresenta as mesmas implicações para tempos entre atendimentos nos modelos de filas. Suponha, particularmente, que existam vários ($n$) *tipos diferentes* de clientes, porém os tempos entre atendimentos para *cada* tipo (tipo $i$) possuem uma distribuição

exponencial com parâmetro $\alpha_i$ ($i = 1, 2, \ldots, n$). Pela Propriedade 2, o tempo *restante* a partir de um instante especificado até a próxima chegada de um cliente do tipo *i* tem a mesma distribuição. Portanto, façamos que $T_i$ seja o tempo restante, medido a partir do instante que um cliente de *qualquer* tipo chegue. A Propriedade 3 nos revela então que *U*, os tempos entre atendimentos para o sistema de filas como um todo, tem uma distribuição exponencial com parâmetro $\alpha$ definido pela última equação. Conseqüentemente, podemos optar por ignorar a distinção entre clientes e ainda ter tempos entre atendimentos exponenciais para o modelo de filas.

Entretanto, as implicações são até mais importantes para *tempos de atendimento* em modelos de filas com vários atendentes que para tempos entre atendimentos. Consideremos, por exemplo, uma situação na qual todos os atendentes possuem a mesma distribuição exponencial de tempo de atendimento com parâmetro $\mu$. Para esse caso, façamos que *n* seja o número de atendentes atendendo *no momento* e façamos que $T_i$ seja o tempo de atendimento *remanescente* para o atendente *i* ($i = 1, 2, \ldots, n$), que também possui uma distribuição exponencial com parâmetro $\alpha_i = \mu$. Decorre então que *U*, o tempo até o término do *próximo* atendimento de qualquer um desses atendentes, tenha uma distribuição exponencial com parâmetro $\alpha = n\mu$. De fato, o sistema de filas *no momento* está funcionando exatamente como um sistema com um *único* atendente no qual tempos de atendimento têm uma distribuição exponencial com parâmetro $n\mu$. Iremos fazer uso freqüente dessa implicação para analisar modelos com vários atendentes posteriormente, ainda no presente capítulo.

Ao usar essa propriedade, algumas vezes também é útil determinar as probabilidades para *quais* das variáveis aleatórias exponenciais por acaso será aquela que tem o valor mínimo. Você poderia, por exemplo, querer encontrar a probabilidade de que determinado atendente *j* terminará de atender um cliente primeiro entre *n* atendentes exponenciais ocupados. É bastante simples (ver Problema 17.4-9) demonstrar que essa probabilidade é proporcional ao parâmetro $\alpha_j$. Particularmente, a probabilidade de que $T_j$ acabará sendo a menor das *n* variáveis aleatórias é

$$P\{T_j = U\} = \alpha_j / \sum_{i=1}^{n} \alpha_i, \qquad \text{para } j = 1, 2, \ldots, n.$$

**Propriedade 4:** Relação com a distribuição de Poisson.

Suponha que o *tempo* entre ocorrências consecutivas de algum tipo particular de evento (por exemplo, chegadas ou términos de atendimento por parte de um atendente permanentemente ocupado) tenha uma distribuição exponencial com parâmetro $\alpha$. A Propriedade 4 tem, então, a ver com a implicação resultante sobre a distribuição probabilística do *número* de vezes que esse tipo de evento ocorre ao longo do tempo especificado. Particularmente, façamos que $X(t)$ seja o número de ocorrências no instante *t* ($t \geq 0$), em que tempo 0 designa o instante no qual começa a contagem. A implicação é que

$$P\{X(t) = n\} = \frac{(\alpha t)^n e^{-\alpha t}}{n!}, \qquad \text{para } n = 0, 1, 2, \ldots;$$

isto é, $X(t)$ possui uma distribuição de Poisson com parâmetro $\alpha t$. Por exemplo, com $n = 0$,

$$P\{X(t) = 0\} = e^{-\alpha t},$$

que é simplesmente a probabilidade da distribuição exponencial de que o *primeiro* evento ocorra após o tempo *t*. A média dessa distribuição de Poisson é

$$E\{X(t)\} = \alpha t,$$

de modo que o número esperado de eventos *por unidade de tempo* seja $\alpha$. Portanto, diz-se que $\alpha$ é a *taxa média* na qual ocorrem os eventos. Quando os eventos são contados de uma forma contínua, diz-se que o processo de contagem $\{X(t); t \geq 0\}$ é um **processo de Poisson** com parâmetro $\alpha$ (a taxa média).

Essa propriedade fornece informações úteis sobre *términos de atendimento* quando tempos de atendimento têm uma distribuição exponencial com parâmetro $\mu$. Obtemos essa informação definindo $X(t)$ como o número de términos de atendimento alcançado por um atendente *permanentemente ocupado* no tempo decorrido $t$, em que $\alpha = \mu$. Para *modelos com vários atendentes* de filas, $X(t)$ também pode ser definido como o número de términos de atendimento alcançado por $n$ atendentes permanentemente ocupados no tempo decorrido $t$, em que $\alpha = n\mu$.

A propriedade é particularmente útil para descrever o comportamento probabilístico das *chegadas* quando tempos entre chegadas possuem uma distribuição exponencial com parâmetro $\lambda$. Nesse caso, $X(t)$ é o *número de* chegadas no tempo decorrido $t$, em que $\alpha = \lambda$ é a *taxa média de chegada.* Portanto, as chegadas ocorrem de acordo com um **processo de entrada de Poisson** com parâmetro $\lambda$. Tais modelos de filas também são descritos como supondo uma *entrada de Poisson.*

Diz-se que as chegadas algumas vezes ocorrem *aleatoriamente,* significando que elas ocorrem de acordo com um processo de entrada de Poisson. Uma interpretação intuitiva desse fenômeno é que todo período de duração fixa tem a *mesma* chance de ter uma chegada independentemente de quando ocorreu a chegada precedente, conforme sugerido pela seguinte propriedade.

**Propriedade 5:** Para todos os valores positivos de $t$, $P\{T \leq t + \Delta t \,|\, T > t\} \approx \alpha \Delta t$, para $\Delta t$ pequeno.

Continuando a interpretar $T$ como o tempo a partir do último evento de certo tipo (chegada ou término de atendimento) até o próximo evento desse tipo, supomos que um tempo $t$ já tenha decorrido sem a ocorrência do evento. Sabemos da Propriedade 2 que a probabilidade de que o evento vá ocorrer dentro do próximo intervalo de tempo de duração fixa $\Delta t$ é uma *constante* (identificada no próximo parágrafo), independentemente de quão grande ou pequeno seja $t$. A Propriedade 5 vai mais além dizendo que, quando o valor de $\Delta t$ é pequeno, essa probabilidade constante pode ser aproximada com boa margem de aproximação por $\alpha \Delta t$. Além disso, ao considerarmos diferentes valores pequenos de $\Delta t$, essa probabilidade é basicamente *proporcional* a $\Delta t$, com fator de proporcionalidade $\alpha$. De fato, $\alpha$ é a *taxa média* na qual ocorrem os eventos (ver Propriedade 4), de modo que o *número esperado* de eventos no intervalo $\Delta t$ seja *exatamente* $\alpha \Delta t$. A única razão para que a probabilidade da ocorrência de um evento vá diferir ligeiramente desse valor é a possibilidade de que ocorra *mais de um* evento, que tem uma probabilidade desprezível quando $\Delta t$ é pequeno.

Para verificar por que a Propriedade 5 é válida matematicamente, note que o valor constante de nossa probabilidade (para um valor fixo $\Delta t > 0$) é simplesmente

$$\begin{aligned} P\{T \leq t + \Delta t \,|\, T > t\} &= P\{T \leq \Delta t\} \\ &= 1 - e^{-\alpha \Delta t}, \end{aligned}$$

para qualquer $t \geq 0$. Portanto, pelo fato de a expansão da série de $e^x$ para qualquer expoente $x$ ser

$$e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!},$$

decorre que

$$P\{T \leq t + \Delta t \,|\, T > t\} = 1 - 1 + \alpha\,\Delta t - \sum_{n=2}^{\infty} \frac{(-\alpha\,\Delta t)^n}{n!}$$
$$\approx \alpha\,\Delta t, \qquad \text{para } \Delta t^5 \text{ pequeno,}$$

pois os termos do somatório se tornam relativamente desprezíveis para valores $\alpha\,\Delta t$ suficientemente pequenos.

Como $T$ pode representar tanto tempos de atendimento como entre chegadas em modelos de filas, essa propriedade fornece uma aproximação conveniente da probabilidade de que o evento de interesse ocorra no próximo intervalo de tempo ($\Delta t$) pequeno. Uma análise baseada nessa aproximação também pode se tornar exata adotando-se os limites apropriados como $\Delta t \to 0$.

**Propriedade 6:** Não é afetada por agregação ou desagregação.

Essa propriedade é relevante basicamente para verificar que o *processo de entrada* é de *Poisson.* Portanto, iremos descrevê-la nesses termos, embora ela também se aplique diretamente à distribuição exponencial (tempos entre atendimentos exponenciais) em virtude da Propriedade 4.

Consideremos primeiramente a agregação (combinada) de diversos processos de entrada de Poisson em um único processo de entrada geral. Particularmente, suponhamos que existam vários ($n$) tipos *diferentes* de clientes, em que os clientes de cada tipo (tipo $i$) cheguem de acordo com um *processo de entrada de Poisson* com parâmetro $\lambda_i$ ($i = 1, 2, \ldots, n$). Supondo que estes sejam processos de Poisson *independentes*, a propriedade diz que o *processo de entrada agregado* (chegada de todos os clientes independentemente do tipo) também deve ser de Poisson, com parâmetro (taxa de chegada) $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$. Em outras palavras, ter um processo de Poisson é *não ser afetado por agregação.*

Essa parte da propriedade decorre diretamente das Propriedades 3 e 4. A última propriedade implica que os tempos entre atendimentos para clientes do tipo $i$ possuem uma distribuição exponencial com parâmetro $\lambda_i$. Para essa mesma situação, já vimos para a Propriedade 3 que ela implica que os tempos entre atendimentos para todos os clientes também têm de ter uma distribuição exponencial, com parâmetro $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$. Usando a Propriedade 4 novamente implica então que o processo de entrada agregado seja de Poisson.

A segunda parte da Propriedade 6 ("não ser afetado por desagregação") refere-se ao caso inverso, no qual o *processo de entrada agregado* (aquele obtido pela combinação de processos de entrada para vários tipos de clientes) é conhecido como Poisson com parâmetro $\lambda$, porém a questão agora se refere à natureza dos processos de entrada *desagregados* (os processos de entrada individuais para os tipos de clientes individuais). Supondo que cada cliente que chega tenha uma probabilidade $p_i$ *fixa* de ser do tipo $i$ ($i = 1, 2, \ldots, n$), com

$$\lambda_i = p_i\lambda \qquad \text{e} \qquad \sum_{i=1}^{n} p_i = 1,$$

a propriedade diz que o processo de entrada para clientes do tipo $i$ também deva ser de Poisson com parâmetro $\lambda_i$. Em outras palavras, ter um processo de Poisson é *não ser afetado por desagregação.*

Como exemplo da utilidade dessa segunda parte da propriedade, considere a seguinte situação. Clientes indistinguíveis chegam de acordo com um processo de Poisson com parâmetro $\lambda$. Cada cliente que chega tem uma probabilidade fixa $p$ de *recusar* (sair sem ter entrado no sistema de filas), de modo que a probabilidade de entrar no sistema seja 1

[5] Mais precisamente,

$$\lim_{\Delta t \to 0} \frac{P\{T \leq t + \Delta t \,|\, T > t\}}{\Delta t} = \alpha.$$

$- p$. Portanto, há dois tipos de clientes — aqueles que se recusam a entrar e aqueles que entram no sistema. A propriedade diz que cada tipo chega de acordo com um processo de Poisson, com parâmetros $p\lambda$ e $(1 - p)\lambda$, respectivamente. Assim, utilizando o último processo de Poisson, modelos de filas que supõem um processo de entrada de Poisson ainda podem ser usados para analisar o desempenho do sistema de filas para aqueles clientes que entram no sistema.

Um dos exemplos na seção de Exemplos Trabalhados do CD-ROM ilustra a aplicação de várias das propriedades da distribuição exponencial apresentada nesta seção.

## 17.5 PROCESSO DE NASCIMENTO-E-MORTE

Os modelos de filas mais elementares partem do pressuposto de que as entradas (clientes que chegam) e saídas (clientes que saem) do sistema de filas ocorram de acordo com o *processo de nascimento-e-morte.* Esse importante processo na teoria das probabilidades tem aplicações em diversas áreas. Entretanto, no contexto da teoria das filas, o termo **nascimento** corresponde à *chegada* de um novo cliente no sistema de filas e a **morte** refere-se à *partida* de um cliente atendido. O *estado* do sistema no instante $t$ ($t \geq 0$), representado por $N(t)$, é o número de clientes no sistema de filas no instante $t$. O processo de nascimento-e-morte descreve *probabilisticamente* como $N(t)$ muda à medida que $t$ aumenta. Em termos genéricos, ela diz que nascimentos e mortes *individuais* ocorrem *aleatoriamente*, em que suas taxas médias de ocorrência dependem apenas do estado atual do sistema. Mais precisamente, as hipóteses do processo de nascimento-e-morte são as seguintes:

**Hipótese 1.** Dado $N(t) = n$, a distribuição probabilística atual do tempo *remanescente* até o próximo *nascimento* (chegada) é *exponencial* com parâmetro $\lambda_n$ ($n = 0, 1, 2, \ldots$).

**Hipótese 2.** Dado $N(t) = n$, a distribuição probabilística atual do tempo *remanescente* até a próxima *morte* (término do atendimento) é *exponencial* com parâmetro $\mu_n$ ($n = 1, 2, \ldots$).

**Hipótese 3.** A variável aleatória da hipótese 1 (o tempo remanescente até o próximo *nascimento*) e a variável aleatória da hipótese 2 (o tempo remanescente até a próxima *morte*) são mutuamente independentes. A próxima transição no estado do processo é

$n \rightarrow n + 1$ (um único nascimento)

ou então

$n \rightarrow n - 1$ (uma única morte),

dependendo de se a primeira ou a última variável aleatória for menor.

Para um sistema de filas, $\lambda_n$ e $\mu_n$ representam, respectivamente, a *taxa média de chegada* e a *taxa média de términos de atendimento,* quando existem $n$ clientes no sistema. Para alguns sistemas de filas, os valores de $\lambda_n$ serão os mesmos para todos os valores de $n$ e os $\mu_n$ também serão os mesmos para todos os $n$, exceto para $n$ muito pequeno (por exemplo, $n = 0$) que um atendente se encontra ocioso. Entretanto, $\lambda_n$ e $\mu_n$ também podem variar consideravelmente com $n$ para alguns sistemas de filas.

Por exemplo, uma das maneiras nas quais $\lambda_n$ pode diferir para diferentes valores de $n$ é o caso no qual vai ficando cada vez mais provável que os possíveis clientes que chegam vão *se recusar* (recusar-se a entrar no sistema) à medida que $n$ aumenta. Da mesma forma, $\mu_n$ pode diferir para $n$ diferentes porque fica cada vez mais provável que os clientes na fila venham a *desistir* (sair sem serem atendidos) à medida que o tamanho da fila aumenta. Um dos exemplos na seção de Exemplos Trabalhados do CD-ROM ilustra um sistema de filas em que ocorrem tanto a recusa quanto a desistência. Esse exemplo demonstra então como os resultados gerais para o processo de nascimento-e-morte conduzem diretamente a várias medidas de desempenho para esse sistema de filas.

### Análise do Processo de Nascimento-e-Morte

Em virtude das suas hipóteses, o processo de nascimento-e-morte é um tipo especial de *cadeia de Markov de tempo contínuo.* Ver Seção 16.8 para uma descrição das cadeias de Markov de tempo contínuo e suas propriedades, inclusive uma introdução ao procedimento geral para encontrar as probabilidades de estado estável que serão aplicadas no restante desta seção. Os modelos de filas que podem ser representados por uma cadeia de Markov de tempo contínuo estão longe de ser mais tratáveis analiticamente que qualquer outro.

Em razão de a Propriedade 4 para a distribuição exponencial (ver Seção 17.4) implicar que $\lambda_n$ e $\mu_n$ são taxas médias, podemos sintetizar essas hipóteses por meio do diagrama de taxas mostrado na Figura 17.4. As setas nesse diagrama mostram as únicas *transições* possíveis para o estado do sistema (conforme especificado pela hipótese 3) e a entrada para cada seta fornece a taxa média para essa transição (conforme especificado pelas hipóteses 1 e 2) quando o sistema se encontra no estado na base da seta.

Exceto para poucos casos especiais, a análise do processo de nascimento-e-morte é muito difícil quando o sistema se encontra em uma condição *transitória*. Alguns resultados sobre a distribuição probabilística de $N(t)$ foram obtidos,[6] mas eles são muito complicados para ser de uso prático. No entanto, é relativamente simples obter essa distribuição *após* o sistema ter atingido uma *condição de estado estável* (supondo que essa condição possa ser alcançada). Essa obtenção pode ser feita diretamente do diagrama de taxas, conforme descrito a seguir.

Considere determinado estado do sistema $n$ ($n = 0, 1, 2, \ldots$). Iniciando no instante 0, suponha que seja feita uma contagem do número de vezes em que o processo entra nesse estado e o número de vezes em que ele deixa esse estado, conforme representado a seguir:

$E_n(t)$ = número de vezes em que o processo entra no estado $n$ no instante $t$.

$L_n(t)$ = número de vezes em que o processo sai do estado $n$ no instante $t$.

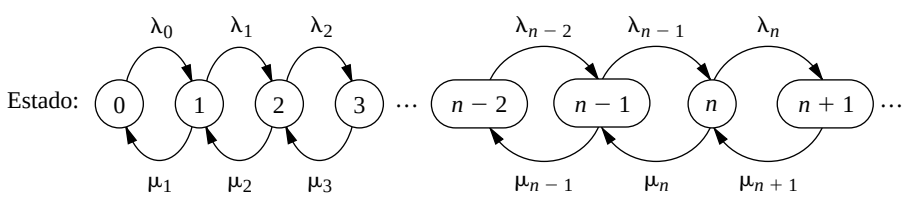
Como os dois tipos de eventos (entrada e saída) têm de ser alternados, esses dois números sempre são iguais ou então diferem de uma unidade, isto é,

$$|E_n(t) - L_n(t)| \leq 1.$$

Dividindo ambos os lados da equação por $t$ e depois fazendo que $t \to \infty$ resulta em

$$\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| \leq \frac{1}{t}, \qquad \text{portanto} \qquad \lim_{t\to\infty}\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| = 0.$$

Dividindo $E_n(t)$ e $L_n(t)$ por $t$ fornece a *taxa real* (número de eventos por unidade de tempo) na qual esses dois tipos de eventos ocorreram e fazendo que $t \to \infty$ dá então a *taxa média* (número esperado de eventos por unidade de tempo):




■ FIGURA 17.4
Diagrama de taxas para o processo de nascimento-e-morte.

[6] KARLIN, S.; MCGREGOR, J. Many Server Queueing Processes with Poisson Input and Exponential Service Times. *Pacific Journal of Mathematics*, v. 8, p. 87-118, 1958.

$$\lim_{t\to\infty} \frac{E_n(t)}{t} = \text{taxa média na qual o processo entra no estado } n.$$

$$\lim_{t\to\infty} \frac{L_n(t)}{t} = \text{taxa média na qual o processo sai do estado } n.$$

Esses resultados conduzem ao seguinte princípio básico:

**Princípio da Taxa que Entra = Taxa que Sai.** Para qualquer estado do sistema $n$ ($n = 0, 1, 2, \ldots$), a taxa média de entrada = taxa média de saída.

A equação expressando esse princípio se chama **equação de equilíbrio** para o estado $n$. Após construir as equações de equilíbrio para todos os estados em termos das probabilidades $P_n$ *desconhecidas*, podemos resolver esse sistema de equações (além de uma equação afirmando que as probabilidades devem somar 1) para encontrar essas probabilidades.

Para ilustrar uma equação de equilíbrio, considere o estado 0. O processo entra nesse estado *somente* a partir do estado 1. Portanto, a probabilidade de estado estável de se encontrar no estado 1 ($P_1$) representa a proporção de tempo que seria *possível* para o processo entrar no estado 0. Dado o processo se encontrar no estado 1, a taxa média de entrada no estado 0 é $\mu_1$. Em outras palavras, para cada unidade de tempo cumulativa que o processo gasta no estado 1, o número esperado de vezes que ele deixaria o estado 1 para entrar no estado 0 é $\mu_1$. De qualquer *outro* estado, essa taxa média é 0. Dessa forma, a taxa média global na qual o processo deixa seu estado atual para entrar no estado 0 (a *taxa média de entrada*) é

$$\mu_1 P_1 + 0(1 - P_1) = \mu_1 P_1.$$

Seguindo o mesmo raciocínio, a *taxa média de saída* tem de ser $\lambda_0 P_0$, de modo que a equação de equilíbrio para o estado 0 é

$$\mu_1 P_1 = \lambda_0 P_0.$$

Para todos os outros estados existem duas transições, ambas entrando e saindo do estado. Portanto, cada lado das equações de equilíbrio para esses estados representa a *soma* das taxas médias para as duas transições envolvidas. Caso contrário, o raciocínio é exatamente o mesmo para o estado 0. Essas equações de equilíbrio são sintetizadas na Tabela 17.1.

Note que a primeira equação de equilíbrio contém duas variáveis a serem resolvidas ($P_0$ e $P_1$), as duas primeiras equações contêm três variáveis ($P_0$, $P_1$ e $P_2$) e assim por diante, de modo que sempre haja uma variável "extra". Por conseguinte, o procedimento para solucionar essas equações é resolver em termos de uma das variáveis, sendo a mais conveniente $P_0$. Por isso, a primeira equação é usada para encontrar $P_1$ em termos de $P_0$; esse resultado e a segunda equação são então usados para encontrar $P_2$ em termos de $P_0$; e assim por diante. No final, a exigência de que a soma de todas as probabilidades seja igual a 1 pode ser usada para calcular $P_0$.

### Resultados para o Processo de Nascimento-e-Morte

Aplicar esse procedimento leva aos seguintes resultados:

■ TABELA 17.1 **Equações de equilíbrio para o processo de nascimento-e-morte**

| Estado | Taxa que Entra = Taxa que Sai |
|---|---|
| 0 | $\mu_1 P_1 = \lambda_0 P_0$ |
| 1 | $\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$ |
| 2 | $\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$ |
| ⋮ | ⋮ |
| $n-1$ | $\lambda_{n-2} P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$ |
| $n$ | $\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$ |
| ⋮ | ⋮ |

*Estado:*

$$
\begin{array}{lllll}
0: & P_1 = \dfrac{\lambda_0}{\mu_1}P_0 & & & \\
1: & P_2 = \dfrac{\lambda_1}{\mu_2}P_1 + \dfrac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) & = \dfrac{\lambda_1}{\mu_2}P_1 & = \dfrac{\lambda_1\lambda_0}{\mu_2\mu_1}P_0 \\
2: & P_3 = \dfrac{\lambda_2}{\mu_3}P_2 + \dfrac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) & = \dfrac{\lambda_2}{\mu_3}P_2 & = \dfrac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}P_0 \\
\vdots & \vdots & & & \\
n-1: & P_n = \dfrac{\lambda_{n-1}}{\mu_n}P_{n-1} + \dfrac{1}{\mu_n}(\mu_{n-1}P_{n-1} - \lambda_{n-2}P_{n-2}) & = \dfrac{\lambda_{n-1}}{\mu_n}P_{n-1} & = \dfrac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}P_0 \\
n: & P_{n+1} = \dfrac{\lambda_n}{\mu_{n+1}}P_n + \dfrac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1}P_{n-1}) & = \dfrac{\lambda_n}{\mu_{n+1}}P_n & = \dfrac{\lambda_n\lambda_{n-1}\cdots\lambda_0}{\mu_{n+1}\mu_n\cdots\mu_1}P_0 \\
\vdots & \vdots & & &
\end{array}
$$

Para simplificar a notação, façamos que

$$C_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}, \qquad \text{para } n = 1, 2, \ldots,$$

e então definamos $C_n = 1$ para $n = 0$. Portanto, as probabilidades de estado estável são

$$P_n = C_n P_0, \qquad \text{para } n = 0, 1, 2, \ldots.$$

A exigência de que

$$\sum_{n=0}^{\infty} P_n = 1$$

implica que

$$\left(\sum_{n=0}^{\infty} C_n\right)P_0 = 1,$$

de modo que

$$P_0 = \left(\sum_{n=0}^{\infty} C_n\right)^{-1}.$$

Quando um modelo de filas se baseia no processo de nascimento-e-morte, de modo que o estado do sistema $n$ represente o número de clientes no sistema de filas, as medidas de desempenho fundamentais para o sistema de filas ($L$, $L_q$, $W$ e $W_q$) podem ser obtidas imediatamente após calcular os $P_n$ das fórmulas anteriores. As definições de $L$ e $L_q$ dadas na Seção 17.2 especificam que

$$L = \sum_{n=0}^{\infty} nP_n, \qquad L_q = \sum_{n=s}^{\infty} (n-s)P_n.$$

Além disso, as relações dadas no final da Seção 17.2 levam a

$$W = \frac{L}{\bar{\lambda}}, \qquad W_q = \frac{L_q}{\bar{\lambda}},$$

em que $\bar{\lambda}$ é a taxa de chegada *média* a longo prazo. Como $\lambda_n$ é a taxa média de chegada enquanto o sistema se encontra no estado $n$ ($n = 0, 1, 2, \ldots$) e $P_n$ é a proporção de tempo de que o sistema se encontra nesse estado,

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n.$$

Diversas das expressões dadas anteriormente envolvem somatórios com um número de termos infinito. Felizmente, esses somatórios possuem soluções analíticas para um número de interessantes casos especiais[7], conforme visto na próxima seção. Caso contrário, eles podem ser aproximados somando-se um número finito de termos via computador.

Esses resultados de estado estável foram obtidos sob a hipótese de que os parâmetros $\lambda_n$ e $\mu_n$ tenham valores tais que o processo possa realmente *alcançar* a condição de estado estável. Essa hipótese *sempre* é válida se $\lambda_n = 0$ para algum valor de $n$ maior que o estado inicial, de modo que sejam possíveis somente um número de estados finito (aqueles menores que esse $n$). Ela *sempre* é válida quando $\lambda$ e $\mu$ são definidos (ver "Terminologia e Notação" na Seção 17.2) e $\rho = \lambda/(s\mu) < 1$. Ela *não* é válida se $\Sigma_{n=1}^{\infty} C_n = \infty$.

A Seção 17.6 descreve vários modelos de filas que são casos especiais do processo de nascimento-e-morte. Portanto, os resultados de estado estável gerais que acabamos de dar nos retângulos serão usados repetidamente para obter resultados de estado estável específicos para esses modelos.

## 17.6 MODELOS DE FILAS BASEADOS NO PROCESSO DE NASCIMENTO-E-MORTE

Como cada uma das taxas médias $\lambda_0, \lambda_1, \ldots$ e $\mu_1, \mu_2, \ldots$ para o processo de nascimento-e-morte pode receber qualquer valor não-negativo, temos grande flexibilidade na modelagem de um sistema de filas. Provavelmente os modelos mais usados na teoria das filas se baseiam diretamente nesse processo. Em virtude das hipóteses 1 e 2 (e a Propriedade 4 para a distribuição exponencial), diz-se que esses modelos possuem uma **entrada de Poisson** e **tempos de atendimento exponenciais**. Os modelos diferem somente em suas hipóteses sobre como $\lambda_n$ e $\mu_n$ mudam com $n$. Apresentamos três desses modelos nesta seção para três tipos importantes dos sistemas de filas.

### Modelo *M/M/s*

Conforme descrito na Seção 17.2, o modelo *M/M/s* parte do pressuposto de que todos os *tempos entre atendimentos* sejam distribuídos de forma independente e idêntica de acordo com uma distribuição exponencial (isto é, o processo de entrada é de Poisson), que todos os *tempos de atendimento* sejam distribuídos de forma independente e idêntica de acordo com outra distribuição exponencial e que o número de atendentes seja $s$ (qualquer inteiro positivo). Conseqüentemente, esse modelo é simplesmente o caso especial do processo de nascimento-e-morte em que a *taxa média de chegada* e a *taxa média de atendimento por atendente ocupado* do sistema de filas são constantes ($\lambda$ e $\mu$, respectivamente) independente do estado do sistema. Quando o sistema tem apenas um *único atendente* ($s = 1$), a implicação é que os parâmetros para o processo de nascimento-e-morte são $\lambda_n = \lambda$ ($n = 0, 1, 2, \ldots$) e $\mu_n = \mu$ ($n = 1, 2, \ldots$). O diagrama de taxas resultante é mostrado na Figura 17.5*a*.

[7] Essas soluções se baseiam nos seguintes resultados conhecidos para a soma de qualquer série geométrica:

$$\sum_{n=0}^{N} x^n = \frac{1 - x^{N+1}}{1 - x}, \qquad \text{para qualquer } x \neq 1,$$

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1 - x}, \qquad \text{se } |x| < 1.$$

Entretanto, quando o sistema tem *vários atendentes* ($s > 1$), $\mu_n$ não pode ser expresso dessa forma tão simples. Tenha em mente que $\mu_n$ representa a taxa média de términos de atendimento para o sistema de filas *global* quando existem $n$ clientes atualmente no sistema. Quando a taxa média de atendimento por atendente ocupado for $\mu$, a taxa média de términos de atendimento global para $n$ atendentes ocupados deve ser $n\mu$. Portanto, $\mu_n = n\mu$ quando $n \leq s$, ao passo que $\mu_n = s\mu$ quando $n \geq s$ de modo que todos os $s$ atendentes estejam ocupados. O diagrama de taxas para esse caso é mostrado na Figura 17.5*b*.

Quando $s\mu$ excede a taxa média de chegada $\lambda$, isto é, quando

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

um sistema de filas que se ajusta a esse modelo vai, finalmente, atingir uma condição de estado estável. Nessa situação, os resultados de estado estável obtidos na Seção 17.5 para o processo de nascimento-e-morte geral são diretamente aplicáveis. Entretanto, esses resultados simplificam consideravelmente para esse modelo e levam a expressões de forma fechada para $P_n$, $L$, $L_q$ e assim por diante, conforme mostrado a seguir.

**Resultados para o Caso com um Único Atendente (*M*/*M*/1).** Para $s = 1$, os fatores $C_n$ para o processo de nascimento-e-morte se reduz a

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n, \qquad \text{para } n = 0, 1, 2, \ldots$$

Portanto,

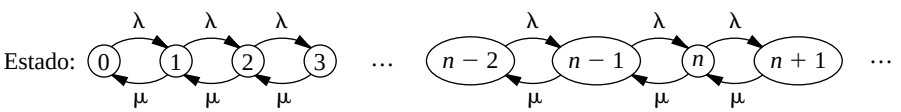
$$P_n = \rho^n P_0, \qquad \text{para } n = 0, 1, 2, \ldots,$$

em que

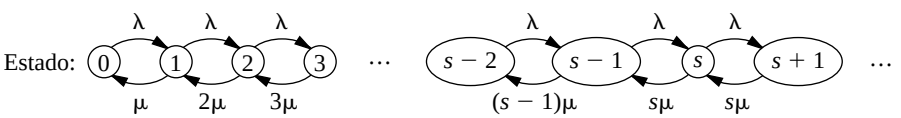
$$P_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1}$$
$$= \left(\frac{1}{1-\rho}\right)^{-1}$$
$$= 1 - \rho.$$

■ FIGURA 17.5
Diagramas de taxas para o modelo *M*/*M*/*s*.

(*a*) *Caso com um único atendente* ($s = 1$) $\lambda_n = \lambda$, para $n = 0, 1, 2, \ldots$; $\mu_n = \mu$, para $n = 1, 2, \ldots$




(*b*) *Caso com vários atendentes* ($s > 1$) $\lambda_n = \lambda$, para $n = 0, 1, 2, \ldots$; $\mu_n = \begin{cases} n\mu, & \text{para } n = 1, 2, \ldots, s \\ s\mu, & \text{para } n = s, s+1, \ldots \end{cases}$



Portanto,

$$P_n = (1 - \rho)\rho^n, \qquad \text{para } n = 0, 1, 2, \ldots .$$

Conseqüentemente,

$$\begin{aligned} L &= \sum_{n=0}^{\infty} n(1 - \rho)\rho^n \\ &= (1 - \rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n) \\ &= (1 - \rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right) \\ &= (1 - \rho)\rho \frac{d}{d\rho}\left(\frac{1}{1 - \rho}\right) \\ &= \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}. \end{aligned}$$

De forma similar,

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty} (n - 1)P_n \\ &= L - 1(1 - P_0) \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)}. \end{aligned}$$

Quando $\lambda \geq \mu$, de modo que a taxa média de chegada exceda a taxa média de atendimento, a solução anterior "estoura" (pois o somatório para calcular $P_0$ diverge). Para esse caso, a fila "explodiria" e cresceria sem limites. Se o sistema de filas iniciar operação sem nenhum cliente presente, o atendente poderia ser bem-sucedido suportando os clientes que chegam ao longo de um curto período, mas isso é impossível no longo prazo. Mesmo quando $\lambda = \mu$, o número de clientes *esperado* no sistema de filas cresce lentamente sem limites ao longo do tempo, pois, embora um retorno temporário para nenhum cliente presente sempre é possível, as probabilidades de números imensos de clientes presentes se torna significativamente maior ao longo do tempo.

Supondo novamente que $\lambda < \mu$, agora podemos obter a distribuição probabilística do *tempo de espera no sistema* (portanto, *incluindo* tempo de atendimento) $\mathcal{W}$ para uma chegada aleatória quando a disciplina da fila é aquela na qual os primeiros que chegam serão os primeiros a ser atendidos. Se essa chegada encontrar $n$ clientes já no sistema, então a chegada terá de esperar ao longo dos $n + 1$ tempos de atendimento exponenciais, inclusive o seu próprio. Para o cliente que está sendo atendido no momento, relembre-se da propriedade de falta de memória para a distribuição exponencial discutida na Seção 17.4. Portanto, façamos que $T_1, T_2, \ldots$ sejam as variáveis aleatórias de tempo de atendimento independentes tendo uma distribuição exponencial com parâmetro $\mu$, e façamos que

$$S_{n+1} = T_1 + T_2 + \ldots + T_{n+1}, \qquad \text{para } n = 0, 1, 2, \ldots ,$$

de modo que $S_{n+1}$ representa o tempo de espera *condicional* dado $n$ clientes já no sistema. Conforme discutido na Seção 17.7, $S_{n+1}$ é conhecido por ter uma *distribuição de Erlang.*[8] Em virtude de a probabilidade de que a chegada aleatória vá encontrar $n$ clientes no sistema ser $P_n$, decorre que

$$P\{\mathcal{W} > t\} = \sum_{n=0}^{\infty} P_n P\{S_{n+1} > t\},$$

[8] Fora do âmbito da teoria das filas, essa distribuição é conhecida como *distribuição gama*.

que reduz após manipulação considerável (ver Problema 17.6-16) para

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

A conclusão surpreendente é que $\mathcal{W}$ tem uma distribuição *exponencial* com parâmetro $\mu(1 - \rho)$. Logo,

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

Esses resultados *incluem* tempo de atendimento no tempo de espera. Em alguns contextos (por exemplo, o problema da sala de emergências do Hospital Municipal), o tempo de espera mais relevante é logo antes de o serviço começar. Portanto, considere o *tempo de espera na fila* (assim, *excluindo* o tempo de atendimento) $W_q$ para uma chegada aleatória quando a disciplina da fila for aquela em que os primeiros que chegam serão os primeiros a ser atendidos. Se essa chegada não encontrar nenhum cliente já no sistema, então a chegada será atendida imediatamente, de modo que

$$P\{W_q = 0\} = P_0 = 1 - \rho.$$

Se, ao contrário, essa chegada encontrar $n > 0$ clientes já na fila, então a chegada tem de esperar por $n$ tempos de atendimento exponenciais até que seu atendimento comece, de forma que

$$\begin{aligned} P\{\mathcal{W}_q > t\} &= \sum_{n=1}^{\infty} P_n P\{S_n > t\} \\ &= \sum_{n=1}^{\infty} (1-\rho)\rho^n P\{S_n > t\} \\ &= \rho \sum_{n=0}^{\infty} P_n P\{S_{n+1} > t\} \\ &= \rho P\{\mathcal{W} > t\} \\ &= \rho e^{-\mu(1-\rho)t}, \qquad \text{para } t \geq 0. \end{aligned}$$

Note que *Wq* não tem nenhuma distribuição exponencial, pois $P\{\mathcal{W}_q = 0\} > 0$. Entretanto, a distribuição *condicional* de $\mathcal{W}_q$, dado que $\mathcal{W}_q > 0$, tem uma distribuição exponencial com parâmetro $\mu(1 - \rho)$, assim como $\mathcal{W}$, pois

$$P\{\mathcal{W}_q > t \,|\, \mathcal{W}_q > 0\} = \frac{P\{\mathcal{W}_q > t\}}{P\{\mathcal{W}_q > 0\}} = e^{-\mu(1-\rho)t}, \qquad \text{para } t \geq 0.$$

Obtendo a média da distribuição (incondicional) de $\mathcal{W}_q$ (ou aplicando $L_q = \lambda W_q$ ou então $W_q = W - 1/\mu$),

$$W_q = E(\mathcal{W}_q) = \frac{\lambda}{\mu(\mu - \lambda)}.$$

Caso queira ver um exemplo que aplique o *modelo M/M/1* para determinar que tipo de equipamento de manipulação de materiais uma empresa deveria comprar, existe um na seção de Exemplos Trabalhados do CD-ROM.

**Resultados para o Caso com Vários Atendentes ($s > 1$).** Quando $s > 1$, os fatores $C_n$ ficam

$$C_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} & \text{para } n = 1, 2, \ldots, s \\ \dfrac{(\lambda/\mu)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s} = \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}} & \text{para } n = s, s+1, \ldots . \end{cases}$$

Conseqüentemente, se $\lambda < s\mu$ [de modo que $\rho = \lambda/(s\mu) < 1$], então

$$\begin{aligned} P_0 &= 1\Big/\left[1 + \sum_{n=1}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{\infty} \left(\frac{\lambda}{s\mu}\right)^{n-s}\right] \\ &= 1\Big/\left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\lambda/(s\mu)}\right], \end{aligned}$$

em que o termo $n = 0$ no último somatório resulta no valor correto igual a 1 em decorrência da convenção que $n! = 1$ quando $n = 0$. Esses fatores $C_n$ também resultam em

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & \text{se } 0 \le n \le s \\ \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}} P_0 & \text{se } n \ge s. \end{cases}$$

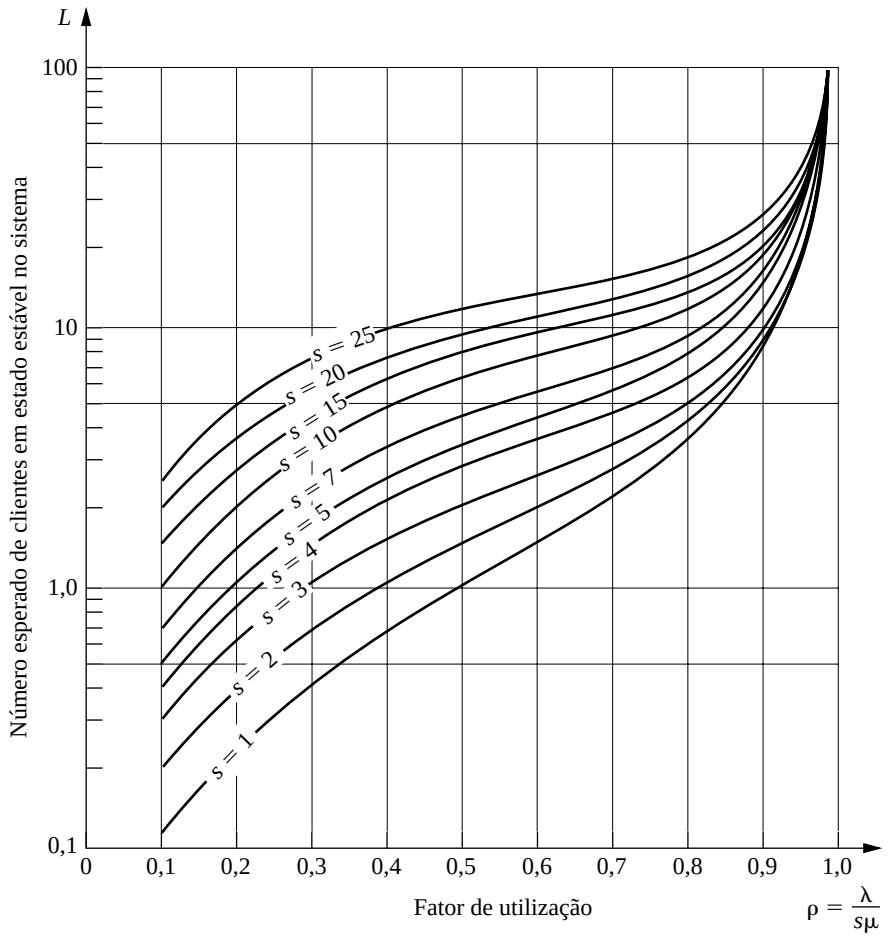
Além disso,

$$\begin{aligned} L_q &= \sum_{n=s}^{\infty} (n-s) P_n \\ &= \sum_{j=0}^{\infty} j P_{s+j} \\ &= \sum_{j=0}^{\infty} j \frac{(\lambda/\mu)^s}{s!} \rho^j P_0 \\ &= P_0 \frac{(\lambda/\mu)^s}{s!} \rho \sum_{j=0}^{\infty} \frac{d}{d\rho}(\rho^j) \\ &= P_0 \frac{(\lambda/\mu)^s}{s!} \rho \frac{d}{d\rho}\left(\sum_{j=0}^{\infty} \rho^j\right) \\ &= P_0 \frac{(\lambda/\mu)^s}{s!} \rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) \\ &= \frac{P_0 (\lambda/\mu)^s \rho}{s!(1-\rho)^2}; \\ W_q &= \frac{L_q}{\lambda}; \\ W &= W_q + \frac{1}{\mu}; \\ L &= \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu}. \end{aligned}$$

A Figura 17.6 mostra como $L$ muda com $\rho$ para vários valores de $s$.

O método com um único atendente para encontrar a distribuição probabilística dos tempos de espera também pode ser estendido para o caso com vários atendentes. Isto resulta[9] (para $t \ge 0$) em

[9] Quando $s - 1 - \lambda/\mu = 0$, $(1 - e^{-\mu t(s-1-\lambda/\mu)})/(s - 1 - \lambda/\mu)$ deveria ser substituído por $\mu t$.




■ FIGURA 17.6
Valores para $L$ para o modelo $M/M/s$ (Seção 17.6).

$$P\{\mathcal{W} > t\} = e^{-\mu t}\left[\frac{1 + P_0(\lambda/\mu)^s}{s!(1-\rho)}\left(\frac{1 - e^{-\mu t(s-1-\lambda/\mu)}}{s - 1 - \lambda/\mu}\right)\right]$$

e

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

em que

$$P\{\mathcal{W}_q = 0\} = \sum_{n=0}^{s-1} P_n.$$

As fórmulas anteriores para as várias medidas de desempenho (inclusive $P_n$) relativamente impõem cálculos manuais. Entretanto, o arquivo em Excel para este capítulo no *Courseware* de PO inclui um gabarito em Excel que realiza todos esses cálculos simultaneamente para quaisquer valores de $t$, $s$, $\lambda$ e $\mu$ que você queira, desde que $\lambda < s\mu$.

Se $\lambda \geq s\mu$, de modo que a taxa média de chegada exceda a taxa média de términos de atendimento máxima, então a fila cresce sem limites, de maneira que as soluções de estado estável anteriores não se aplicam.

**Exemplo do Hospital Municipal com o Modelo *M/M/s*.** Para o problema da sala de emergências do Hospital Municipal (ver Seção 17.1), o administrador concluiu que os casos de emergência chegam, em sua maioria, de forma aleatória (um *processo de entrada de Poisson*), de modo que os tempos entre atendimentos possuem uma distribuição exponencial. Ele também concluiu que o tempo gasto por um médico tratando os casos segue, aproximadamente, uma *distribuição exponencial*. Assim, ele optou pelo modelo *M/M/s* para um estudo preliminar desse sistema de filas.

Projetando para o ano que vem os dados disponíveis para o turno do início da noite, ele estima que os pacientes chegarão em uma taxa *média* de 1 a cada 1/2 hora. Um médico precisa em média de 20 minutos para atender cada paciente. Portanto, com uma hora sendo a unidade de tempo,

$$\frac{1}{\lambda} = \frac{1}{2} \text{ hora por cliente}$$

e

$$\frac{1}{\mu} = \frac{1}{3} \text{ hora por cliente,}$$

de modo que

$$\lambda = 2 \text{ clientes por hora}$$

e

$$\mu = 3 \text{ clientes por hora.}$$

As duas alternativas consideradas são para continuar a ter apenas um médico durante esse turno ($s = 1$) ou então disponibilizar um segundo médico ($s = 2$). Em ambos os casos,

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

de forma que o sistema deveria aproximar-se de uma condição de estado estável. Na verdade, como $\lambda$ é ligeiramente distinto durante outros turnos, o sistema jamais atingirá realmente uma condição de estado estável, porém o administrador acha que os resultados de estado estável fornecerão uma boa aproximação. Portanto, as equações anteriores são usadas para obter os resultados mostrados na Tabela 17.2.

Com base nesses resultados, ele concluiu provisoriamente que um único médico seria inadequado para o próximo ano para fornecer os cuidados relativamente imediatos necessários para uma sala de emergências de um hospital. Veremos adiante (Seção 17.8) como ele chegou a essa conclusão aplicando outro modelo de filas que fornece uma representação melhor do real sistema de filas de maneira crucial.

Para mais um exemplo da aplicação do modelo *M/M/1* consulte a seção de Exemplos Trabalhados do CD-ROM, em que a questão nesse caso é se três empregados em uma lanchonete deveriam trabalhar juntos funcionando como um único atendente rápido ou então separadamente como três atendentes consideravelmente lentos.

■ TABELA 17.2 **Resultados de estado estável do modelo *M/M/s* para o problema do Hospital Municipal**

| | $s = 1$ | $s = 2$ |
|---|---|---|
| $\rho$ | $\frac{2}{3}$ | $\frac{1}{3}$ |
| $P_0$ | $\frac{1}{3}$ | $\frac{1}{2}$ |
| $P_1$ | $\frac{2}{9}$ | $\frac{1}{3}$ |
| $P_n$ para $n \geq 2$ | $\frac{1}{3}\left(\frac{2}{3}\right)^n$ | $\left(\frac{1}{3}\right)^n$ |
| $L_q$ | $\frac{4}{3}$ | $\frac{1}{12}$ |
| $L$ | 2 | $\frac{3}{4}$ |
| $W_q$ | $\frac{2}{3}$ hora | $\frac{1}{24}$ hora |
| $W$ | 1 hora | $\frac{3}{8}$ hora |
| $P\{W_q > 0\}$ | 0.667 | 0.167 |
| $P\left\{W_q > \frac{1}{2}\right\}$ | 0.404 | 0.022 |
| $P\{W_q > 1\}$ | 0.245 | 0.003 |
| $P\{W_q > t\}$ | $\frac{2}{3}e^{-t}$ | $\frac{1}{6}e^{-4t}$ |
| $P\{W > t\}$ | $e^{-t}$ | $\frac{1}{2}e^{-3t}(3 - e^{-t})$ |

### Variante de Fila Finita do Modelo *M/M/s* (Denominado Modelo *M/M/s/K*)

Mencionamos na discussão sobre filas na Seção 17.2 que os sistemas de filas algumas vezes têm uma *fila finita*, isto é, o número de clientes no sistema não pode ultrapassar algum número especificado (representado por $K$) de modo que a capacidade da fila é $K - s$. Qualquer cliente que chegue enquanto a fila estiver "cheia" não pode entrar no sistema e, portanto, sai para sempre. Do ponto de vista do processo de nascimento-e-morte, a taxa média de entrada no sistema se torna zero nesses momentos. Portanto, a única modificação necessária no *modelo M/M/s* para introduzir uma fila finita é alterar os parâmetros $\lambda_n$ para

$$\lambda_n = \begin{cases} \lambda & \text{para } n = 0, 1, 2, \ldots, K-1 \\ 0 & \text{para } n \geq K. \end{cases}$$

Como $\lambda_n = 0$ para alguns valores de $n$, um sistema de filas que se ajusta a esse modelo sempre vai finalmente atingir uma condição de estado estável, mesmo quando $\rho = \lambda/s\mu \leq 1$.

Esse modelo é chamado comumente *M*/*M*/s/*K*, em que a presença do quarto símbolo o distingue do modelo *M/M/s*. A única diferença na formulação desses dois modelos é que $K$ é finito para o modelo *M*/*M*/s/*K* e $K = \infty$ para o modelo *M/M/s*.

A interpretação física usual para o modelo *M*/*M*/s/*K* é que existe apenas uma *sala de espera limitada* que acomodará um máximo de $K$ clientes no sistema. Por exemplo, para o problema da sala de emergências do Hospital Municipal, esse sistema na verdade teria uma fila finita se houvesse apenas $K$ leitos para os pacientes e se a política fosse a de enviar pacientes que chegam para outro hospital toda vez que não houvesse leitos vazios.

Outra interpretação possível é que os clientes que chegam sairão e "procurarão outro caminho" toda vez que eles encontrarem muitos clientes ($K$) à sua frente no sistema, pois eles não estão propensos a esperar muito nessa fila. Esse fenômeno de recusa é bastante comum

em sistemas de atendimento comerciais. Entretanto, existem outros modelos disponíveis (por exemplo, ver Problema 17.5-5) que se encaixam ainda melhor nessa interpretação.

O diagrama de taxas para esse modelo é idêntico àquele mostrado na Figura 17.5 para o *modelo M/M/s*, *exceto* que ele pára com o estado *K*.

**Resultados para o Caso com um Único Atendente (*M/M/1/K*).** Para esse caso,

$$C_n = \begin{cases} \left(\dfrac{\lambda}{\mu}\right)^n = \rho^n & \text{para } n = 0, 1, 2, \ldots, K \\ 0 & \text{para } n > K. \end{cases}$$

Portanto, para $\rho \neq 1$,[10]

$$\begin{aligned} P_0 &= \frac{1}{\sum_{n=0}^{K} (\lambda/\mu)^n} \\ &= 1 \Big/ \left[ \frac{1 - (\lambda/\mu)^{K+1}}{1 - \lambda/\mu} \right] \\ &= \frac{1 - \rho}{1 - \rho^{K+1}}, \end{aligned}$$

de modo que

$$P_n = \frac{1 - \rho}{1 - \rho^{K+1}} \rho^n, \qquad \text{para } n = 0, 1, 2, \ldots, K.$$

Portanto,

$$\begin{aligned} L &= \sum_{n=0}^{K} nP_n \\ &= \frac{1 - \rho}{1 - \rho^{K+1}} \rho \sum_{n=0}^{K} \frac{d}{d\rho}(\rho^n) \\ &= \frac{1 - \rho}{1 - \rho^{K+1}} \rho \frac{d}{d\rho}\left( \sum_{n=0}^{K} \rho^n \right) \\ &= \frac{1 - \rho}{1 - \rho^{K+1}} \rho \frac{d}{d\rho}\left( \frac{1 - \rho^{K+1}}{1 - \rho} \right) \\ &= \rho \frac{-(K+1)\rho^K + K\rho^{K+1} + 1}{(1 - \rho^{K+1})(1 - \rho)} \\ &= \frac{\rho}{1 - \rho} - \frac{(K+1)\rho^{K+1}}{1 - \rho^{K+1}}. \end{aligned}$$

Como de praxe (quando $s = 1$),

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

Observe que os resultados anteriores não precisam que $\lambda < \mu$ (isto é, que $\rho < 1$).

Quando $\rho < 1$, pode ser verificado que o segundo termo na expressão final para $L$ converge para 0 à medida que $K \to \infty$, de forma que *todos* os resultados anteriores de fato convirjam para os resultados correspondentes ao modelo *M/M/1*, mostrado anteriormente.

As distribuições de tempos de espera podem ser obtidas usando-se o mesmo raciocínio daquele para o modelo *M/M/1* (ver Problema 17.6-27). Entretanto, não são obtidas expressões simples nesse caso, de modo que é necessário o emprego de computador para efetuar

[10] Se $\rho = 1$, então $P_n = 1/(K + 1)$ para $n = 0, 1, 2, \ldots, K$, de modo que $L = K/2$.

os cálculos. Felizmente, embora $L \neq \lambda W$ e $L_q \neq \lambda W_q$ para o modelo atual, pois $\lambda_n$ não é igual para todo $n$ (ver o final da Seção 17.2), os tempos de espera *previstos* para clientes entrando no sistema ainda podem ser obtidos diretamente das expressões dadas no final da Seção 17.5:

$$W = \frac{L}{\bar{\lambda}}, \qquad W_q = \frac{L_q}{\bar{\lambda}},$$

em que

$$\begin{aligned}\bar{\lambda} &= \sum_{n=0}^{\infty} \lambda_n P_n \\ &= \sum_{n=0}^{K-1} \lambda P_n \\ &= \lambda(1 - P_K).\end{aligned}$$

**Resultados para o Caso com Vários Atendentes ($s > 1$).** Como esse modelo não permite mais que $K$ clientes no sistema, $K$ é o número máximo de atendentes que poderia ser usado. Portanto, suponha que $s \leq K$. Nesse caso, $C_n$ se torna

$$C_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} & \text{para } n = 0, 1, 2, \ldots, s \\ \dfrac{(\lambda/\mu)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s} = \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}} & \text{para } n = s, s+1, \ldots, K \\ 0 & \text{para } n > K. \end{cases}$$

Portanto,

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & \text{para } n = 1, 2, \ldots, s \\ \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}} P_0 & \text{para } n = s, s+1, \ldots, K \\ 0 & \text{para } n > K, \end{cases}$$

em que

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

Essas fórmulas continuam a usar a convenção de que $n! = 1$ quando $n = 0$.
Adaptando a derivada de $L_q$ para o modelo *M/M/s* a esse caso resulta em

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}\left[1 - \rho^{K-s} - (K-s)\rho^{K-s}(1-\rho)\right],$$

em que $\rho = \lambda/(s\mu)$.[11] Pode se provar que

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right).$$

[11] Se $\rho = 1$, é necessário aplicar duas vezes a regra de L'Hôpital a essa expressão para $L_q$. Caso contrário, todos esses resultados com vários atendentes serão válidos para todo $\rho > 0$. A razão para que esse sistema de filas consiga atingir uma condição de estado estável mesmo quando $\rho \geq 1$ é que $\lambda_n = 0$ para $n \geq K$, de modo que o número de clientes no sistema não possa continuar a crescer indefinidamente.

E $W$ e $W_q$ são obtidos desses valores exatamente como mostrado para o caso com um único atendente.

O arquivo Excel para este capítulo inclui um gabarito em Excel para calcular as medidas de desempenho dadas anteriormente (inclusive $P_n$) para esse modelo.

Um caso especial interessante desse modelo é aquele no qual $K = s$ de maneira que a capacidade da fila seja $K - s = 0$. Nesse caso, clientes que chegam quando todos os atendentes estão ocupados sairão imediatamente e serão perdidos para o sistema. Isso ocorreria, por exemplo, em uma rede telefônica com $s$ linhas-tronco de modo que aqueles que telefonassem receberiam um sinal de ocupado e desligariam quando todas as linhas-tronco estivessem ocupadas. Esse tipo de sistema (um "sistema de filas" sem nenhuma fila) é chamado *sistema de perda de Erlang*, pois ele foi estudado pela primeira vez no início do século XX por A. K. Erlang, um engenheiro de telecomunicações dinamarquês que é considerado o criador da teoria das filas.

Hoje é comum para o sistema telefônico em um *call center* fornecer algumas linhas-tronco extras que colocam a pessoa que fez a chamada em espera, porém, outras pessoas que ligarem depois disso poderão encontrar as linhas ocupadas (sinal de ocupado). Um sistema destes também se ajusta a esse modelo, no qual $(K - s)$ é o número de linhas-tronco extras que colocam a pessoa que fez a chamada na espera. Um dos exemplos na seção de Exemplos Trabalhados do CD-ROM ilustra a aplicação desse modelo para um sistema destes.

### Variante da População Solicitante Finita do Modelo *M/M/s*

Suponha agora que o único desvio do modelo *M/M/s* seja que (conforme definido na Seção 17.2) a fonte de entradas seja *limitada*, isto é, o tamanho da população solicitante é *finito*. Para esse caso, façamos que $N$ represente o tamanho da população solicitante. Assim, quando o número de clientes no sistema de filas for $n$ ($n = 0, 1, 2, \ldots, N$), existirão apenas $N - n$ *possíveis* clientes restantes na fonte de entradas.

A aplicação mais importante desse modelo foi a do problema do conserto de máquinas, no qual um ou mais técnicos de manutenção recebem o encargo de manter em operação certo grupo de $N$ máquinas reparando cada uma que quebrar. O exemplo dado no final da Seção 16.8 ilustra essa aplicação quando os procedimentos gerais para solucionar qualquer *cadeia de Markov de tempo contínuo* são usados em vez das fórmulas específicas disponíveis para o processo de nascimento-e-morte. A equipe de manutenção é considerada como atendentes individuais no sistema de filas se eles trabalharem individualmente em diferentes tipos de máquinas, ao passo que toda a equipe é considerada um único atendente se os membros da equipe trabalharem juntos em cada máquina. As máquinas constituem a população solicitante. Cada uma delas é considerada um cliente no sistema de filas quando se encontrar quebrada aguardando ser reparada, ao passo que ela se encontra fora do sistema de filas enquanto ela estiver operacional.

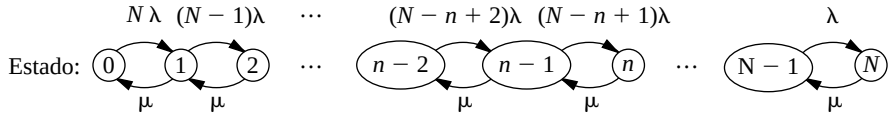
Note que cada membro da população solicitante alterna entre estar *dentro* e *fora* do sistema de filas. Portanto, o análogo do *modelo M/M/s* que se enquadra nessa situação supõe que o *tempo fora* de *cada* membro (isto é, o tempo decorrido entre deixar o sistema até retornar da próxima vez) tem uma distribuição exponencial com parâmetro $\lambda$. Quando $n$ dos membros estiverem *dentro* e, portanto, $N - n$ membros estiverem *fora*, a distribuição probabilística atual do tempo *restante* até a próxima chegada ao sistema de filas é a distribuição do *mínimo* dos *tempos fora restantes* para os últimos $N - n$ membros. As Propriedades 2 e 3 para a distribuição exponencial implicam que essa distribuição tem de ser exponencial com parâmetro $\lambda_n = (N - n)\lambda$. Assim, esse modelo é simplesmente o caso especial do processo de nascimento-e-morte que tem o diagrama de taxas mostrado na Figura 17.7.

Como $\lambda_n = 0$ para $n = N$, qualquer sistema de filas que se ajuste a esse modelo acabará finalmente atingindo uma condição de estado estável. Os resultados de estado estável disponíveis são sintetizados a seguir:

**Resultados para o Caso com um Único Atendente (s = 1).** Quando s = 1, os $C^n$ fatores na Seção 17.5 reduzem-se a

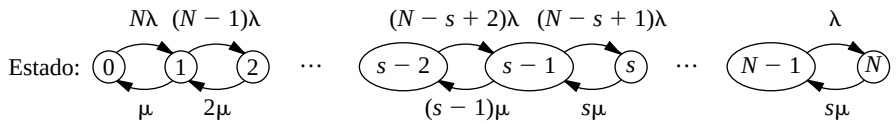
(*a*) *Caso com um único atendente* ($s = 1$)

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{para } n = 0, 1, 2, ..., N \\ 0, & \text{para } n \geq N \end{cases}$$
$$\mu_n = \mu, \quad \text{para } n = 1, 2, ...$$




(*b*) *Caso com vários atendentes* ($s > 1$)

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{para } n = 0, 1, 2, ..., N \\ 0, & \text{para } n \geq N \end{cases}$$
$$\mu_n = \begin{cases} n\mu, & \text{para } n = 1, 2, ..., s \\ s\mu, & \text{para } n = s, s+1, ... \end{cases}$$




■ FIGURA 17.7
Diagramas de taxas para uma variação de população solicitante finita do modelo *M*/*M*/*s*.

$$C_n = \begin{cases} N(N-1)\cdots(N-n+1)\left(\dfrac{\lambda}{\mu}\right)^n = \dfrac{N!}{(N-n)!}\left(\dfrac{\lambda}{\mu}\right)^n & \text{para } n \leq N \\ 0 & \text{para } n > N, \end{cases}$$

para esse modelo. Portanto, usando novamente a convenção de que $n! = 1$ quando $n = 0$,

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

$$P_n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0, \qquad \text{se } n = 1, 2, \ldots, N;$$

$$L_q = \sum_{n=1}^{N} (n-1)P_n,$$

que pode ser reduzida a

$$L_q = N - \frac{\lambda + \mu}{\lambda}(1 - P_0);$$

$$L = \sum_{n=0}^{N} nP_n = L_q + 1 - P_0$$

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

Finalmente,

$$W = \frac{L}{\overline{\lambda}} \qquad \text{e} \qquad W_q = \frac{L_q}{\overline{\lambda}},$$

em que

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

Neste ponto, você poderia achar útil retornar ao exemplo no final da Seção 16.8, pois aquele exemplo se encaixa perfeitamente nesse modelo de caso com um único atendente. Particularmente, $N = 2$, $\lambda = 1$ e $\mu = 2$ para aquele exemplo, de modo que $P_0 = 0{,}4$, $P_1 = 0{,}4$, $P_2 = 0{,}2$ e assim por diante.

**Resultados para o Caso com Vários Atendentes (s > 1).** Para $N \geq s > 1$,

$$C_n = \begin{cases} \dfrac{N!}{(N-n)!\,n!}\left(\dfrac{\lambda}{\mu}\right)^n & \text{para } n = 0, 1, 2, \ldots, s \\ \dfrac{N!}{(N-n)!s!s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n & \text{para } n = s, s+1, \ldots, N \\ 0 & \text{para } n > N. \end{cases}$$

Portanto,

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

em que

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

Finalmente,

$$L_q = \sum_{n=s}^{N} (n-s)P_n$$

e

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right),$$

que então resulta em $W$ e $W_q$ pelas mesmas equações como no caso com um único atendente.

O arquivo Excel para este capítulo inclui um gabarito em Excel para realizar todos os cálculos anteriores.

Estão disponíveis também tabelas extensas de resultados computacionais[12] para esse modelo tanto para o caso com um único atendente como para aquele com vários atendentes.

Para ambos os casos, foi demonstrado[13] que as fórmulas anteriores para $P_n$ e $P_0$ (e, portanto, para $L_q$, $L$, $W$ e $W_q$) *também* são válidas para uma generalização desse modelo. Particularmente, podemos *eliminar* a hipótese de que os tempos gastos *fora* do sistema de filas pelos membros da população solicitante possuam uma distribuição exponencial, embora isso tire o modelo fora do escopo do processo de nascimento-e-morte. Desde que esses tempos sejam distribuídos identicamente com média $1/\lambda$ (e a hipótese dos tempos de atendimento exponenciais ainda é válida), esses tempos fora podem ter *qualquer* distribuição probabilística!

[12] PECK, L. G.; HAZELWOOD, R. N. *Finite Queueing Tables*. Nova York: Wiley, 1958.

[13] BUNDAY, B. D.; SCRATON, R. E. The G/M/r Machine Interference Model. *European Journal of Operational Research*, v. 4, p. 399-402, 1980.

## 17.7 MODELOS DE FILAS ENVOLVENDO DISTRIBUIÇÕES NÃO-EXPONENCIAIS

Como todos os modelos de teoria das filas na seção anterior (exceto para uma generalização) se baseiam no processo de nascimento-e-morte, tanto os tempos entre chegadas quanto os tempos de atendimento precisam ter distribuições exponenciais. Conforme discutido na Seção 17.4, esse tipo de distribuição probabilística possui muitas propriedades convenientes para teoria das filas, mas ele fornece uma adequação razoável para apenas certos tipos de sistemas de filas. Em particular, a hipótese dos tempos entre atendimentos exponenciais implica que as chegadas ocorrem aleatoriamente (um processo de entrada de Poisson), que é uma aproximação razoável em muitas situações, mas *não* para o caso em que as chegadas são cuidadosamente programadas ou reguladas. Além disso, a distribuição de tempo de atendimento real freqüentemente se desvia muito da forma exponencial, particularmente quando as exigências de atendimento dos clientes são bastante similares. Portanto, é importante ter disponível outros modelos de filas que usem distribuições alternativas.

Infelizmente, a análise matemática dos modelos de filas como distribuições não-exponenciais é muito mais difícil. Entretanto, foi possível se obter alguns resultados úteis para alguns desses modelos. Essa análise está fora do escopo deste livro, porém, nesta seção, iremos fazer um resumo dos modelos e descrever seus resultados.

### Modelo *M/G/*1

Conforme introduzido na Seção 17.2, o modelo *M/G/*1 parte do pressuposto de que o sistema de filas tem um *único atendente* e um *processo de entrada de Poisson* (tempos entre atendimentos exponenciais) com uma taxa média de chegada *fixa*, $\lambda$. Como de praxe, supõe-se que os clientes tenham tempos de atendimento *independentes* com a *mesma* distribuição probabilística. Entretanto, não é imposta nenhuma restrição de como deve ser essa distribuição de tempos de atendimento. Na realidade, é necessário apenas conhecer (ou estimar) a média $1/\mu$ e a variância $\sigma^2$ dessa distribuição.

Qualquer sistema de filas desses pode finalmente atingir uma condição de estado estável se $\rho = \lambda/\mu < 1$. Os resultados de estado estável[14] prontamente disponíveis para esse modelo geral são os seguintes:

$$P_0 = 1 - \rho,$$
$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)},$$
$$L = \rho + L_q,$$
$$W_q = \frac{L_q}{\lambda},$$
$$W = W_q + \frac{1}{\mu}.$$

Considerando a complexidade envolvida na análise de um modelo que permita *qualquer* distribuição de tempo de atendimento, é incrível que uma fórmula tão simples possa ser obtida para $L_q$. Essa fórmula é um dos resultados mais importantes na teoria das filas em razão de sua facilidade de uso e o predomínio de sistemas de filas *M*/*G*/1 na prática. Essa equação para $L_q$ (ou seu equivalente para $W_q$) é comumente chamada **fórmula de Pollaczek-Khintchine**, em homenagem aos dois pioneiros no desenvolvimento da teoria das filas que obtiveram a fórmula no início dos anos 30.

[14] Existe também uma fórmula de recursão para calcular a distribuição probabilística do número de clientes no sistema; ver HORDIJK, A.; TIJMS, H. C. A Simple Proof of the Equivalence of Limiting Distribution of the Continuous-Time and the Embedded Process of Queue Size in the *M*/*G*/1 Queue. *Statistica Neerlandica*, v. 36, p. 97-100, 1976.

Para qualquer tempo de atendimento esperado fixo, $1/\mu$, note que $L_q$, $L$, $W_q$ e $W$ aumentam à medida que $\sigma^2$ é aumentado. Esse resultado é importante, pois ele indica que a regularidade do atendente tem uma importante relação com o desempenho da instalação de atendimento — não apenas a velocidade média do atendente. Esse ponto-chave é ilustrado na próxima subseção.

Quando a distribuição de tempos de atendimento for exponencial, $\sigma^2 = 1/\mu^2$, e os resultados anteriores reduzirão os resultados correspondentes para o *modelo M/M/*1 dado no início da Seção 17.6.

A flexibilidade total na distribuição de tempos de atendimento fornecida por esse modelo é extremamente útil, de modo que é uma pena que esforços para obter resultados similares para o caso com vários atendentes tenham sido infrutíferos. Entretanto, alguns resultados com vários atendentes foram obtidos para os importantes casos especiais descritos pelos dois modelos a seguir. Existem gabaritos em Excel no arquivo Excel para este capítulo para realizar os cálculos tanto para o modelo *M/G/*1 como para os dois modelos considerados a seguir quando $s = 1$.

### Modelo *M/D/s*

Quando o atendimento consiste essencialmente na mesma tarefa rotineira a ser realizada para todos os clientes, há uma tendência de haver pouca variação no tempo de atendimento necessário. O modelo *M*/*D*/*s* normalmente fornece uma representação razoável para esse tipo de situação, pois ele supõe que todos os tempos de atendimento se igualam a alguma *constante* fixa (a distribuição *degenerada* de tempo de atendimento) e que temos um *processo de entrada de Poisson* com uma taxa média de chegada fixa $\lambda$.

Quando há apenas um atendente, o modelo *M/D/*1 é simplesmente o caso especial do modelo *M/G/*1 em que $\sigma^2 = 0$, de modo que a *fórmula de Pollaczek-Khintchine* reduz-se a

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

em que $L$, $W_q$ e $W$ são obtidos de $L_q$ conforme ilustrado a seguir. Note que estes $L_q$ e $W_q$ são exatamente *metade* do tamanho daqueles para o caso de tempo de atendimento exponencial da Seção 17.6 (o modelo *M/M/*1), no qual $\sigma^2 = 1/\mu^2$, de modo que diminuir $\sigma^2$ pode melhorar *muito* a medida de desempenho de um sistema de filas.

Para a versão com vários atendentes desse modelo (*M*/*D*/*s*), existe um método complicado[15] para obter a distribuição probabilística de estado estável do número de clientes no sistema e sua média [supondo que $\rho = \lambda/(s\mu) < 1$]. Esses resultados foram tabulados para inúmeros casos,[16] e as médias ($L$) também são dadas graficamente na Figura 17.8.

### Modelo $M/E^k/s$

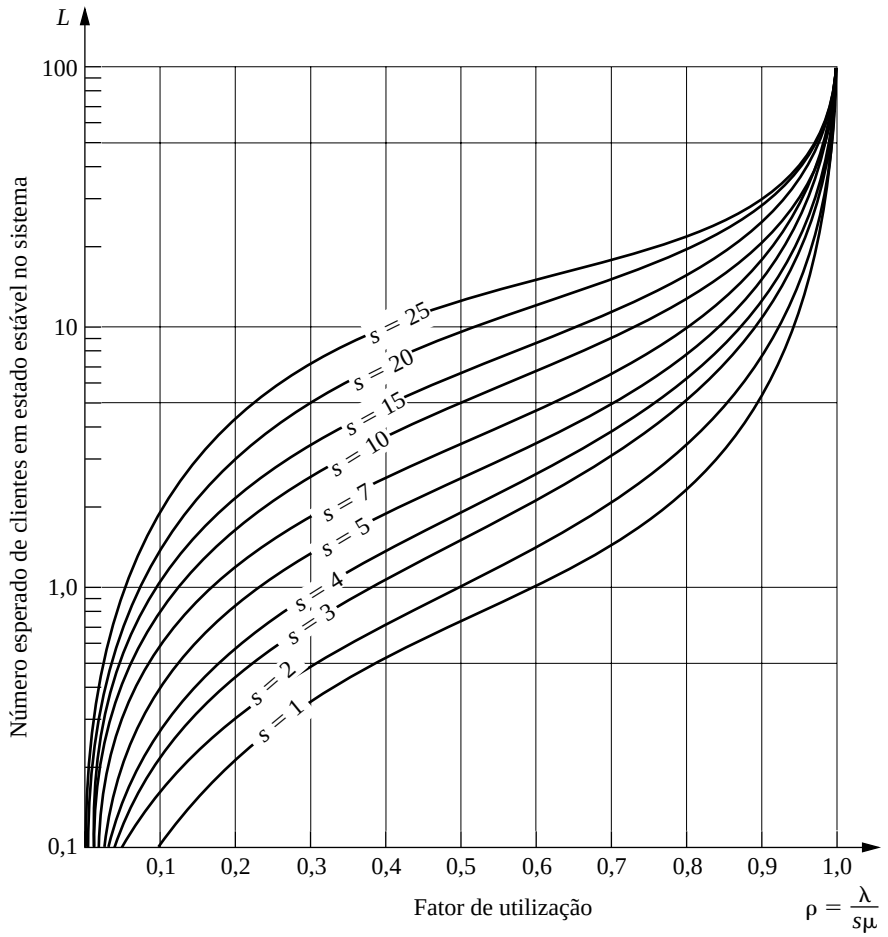
O modelo *M/D/s* supõe uma variação *zero* nos tempos de atendimento ($\sigma = 0$), ao passo que a distribuição exponencial de tempos de atendimento supõe uma variação muito grande ($\sigma = 1/\mu$). Entre esses dois casos bastante extremos temos um intermédio extenso ($0 < \sigma < 1/\mu$), no qual a maioria das distribuições de tempos de atendimento *reais* caem. Outro tipo de distribuição de tempos de atendimento teórica que cai nesse meio-termo é a **distribuição de Erlang** (em homenagem ao criador da teoria das filas).

A função de densidade probabilística para a distribuição de Erlang é

$$f(t) = \frac{(\mu k)^k}{(k-1)!}\, t^{k-1} e^{-k\mu t}, \qquad \text{para } t \geq 0,$$

[15] Ver PRABHU, N. U. *Queues and Inventories*, Nova York. Wiley, p. 32-34, 1965; ver também as páginas 286-288 na Referência Selecionada 5.

[16] HILLIER, F. S. et al. *Queueing Tables and Graphs.* Nova York: Elsevier North-Holland, 1981.




■ FIGURA 17.8
Valores de $L$ para o modelo $M/D/s$ (Seção 17.7).

em que $\mu$ e $k$ são parâmetros da distribuição estritamente positivos e, além disso, $k$ também é restrito a ser inteiro. Exceto por essa restrição inteira e a definição dos parâmetros, essa distribuição é *idêntica* à *distribuição gama*. Sua média e desvio-padrão são

$$\text{Média} = \frac{1}{\mu}$$

e

$$\text{Desvio-padrão} = \frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}} \frac{1}{\mu}.$$

Portanto, $k$ é o parâmetro que especifica o grau de variabilidade dos tempos de atendimento relativos à média. Normalmente ele é conhecido como *parâmetro de forma.*

A distribuição de Erlang é uma distribuição muito importante na teoria das filas por duas razões. Para descrever a primeira, suponha que $T_1, T_2, \ldots, T_k$ sejam $k$ variáveis aleatórias independentes com uma distribuição exponencial idêntica cuja média é $1/(k\mu)$. Sua soma então

$$T = T_1 + T_2 + \ldots + T_k$$

possui uma *distribuição de Erlang* com parâmetros $\mu$ e $k$. A discussão da distribuição exponencial na Seção 17.4 sugeria que o tempo necessário para realizar certos tipos de tarefas poderiam muito bem ter distribuição exponencial. Entretanto, o atendimento total

necessário para um cliente poderia envolver o desempenho do atendente realizando não somente uma tarefa específica, mas sim uma seqüência de $k$ tarefas. Se as respectivas tarefas tiverem uma distribuição exponencial idêntica e independente para suas durações, o tempo de atendimento total terá uma distribuição de Erlang. Este seria o caso, por exemplo, se o atendente tivesse de realizar a *mesma* tarefa exponencial $k$ vezes independentes para cada cliente.

A distribuição de Erlang também é muito útil, pois ela é uma grande (dois parâmetros) família de distribuições permitindo somente valores não-negativos. Assim, distribuições de tempos de atendimento empíricas podem normalmente ser razoavelmente aproximadas por uma distribuição de Erlang. De fato, tanto as distribuições *exponenciais* quanto as *degeneradas* (constantes) são casos especiais da distribuição de Erlang, com $k = 1$ e $k = \infty$, respectivamente. Valores intermediários de $k$ fornecem distribuições intermediárias com média $= 1/\mu$, modo $= (k - 1)/(k\mu)$ e variância $= 1/(k\mu^2)$, conforme sugerido pela Figura 17.9. Portanto, após estimar a média e variância de uma distribuição de tempos de atendimento empírica, essas fórmulas para a média e variância podem ser usadas para escolher o valor inteiro de $k$ que se aproxima mais de perto das estimativas.

Considere agora o modelo $M/E_k/1$, que é simplesmente o caso especial do modelo $M/G/$l, no qual tempos de atendimento possuem uma distribuição de Erlang com parâmetro de forma $= k$. Aplicando a fórmula de Pollaczek-Khintchine com $\sigma^2 = 1/(k\mu^2)$ (e os resultados dados para $M/G/1$) resulta em

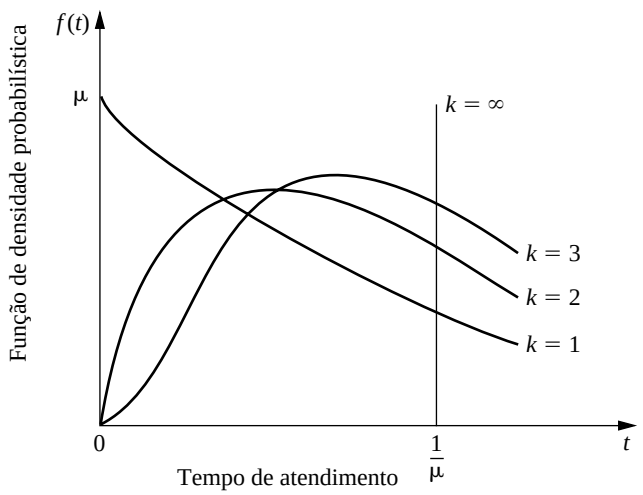
$$L_q = \frac{\lambda^2/(k\mu^2) + \rho^2}{2(1-\rho)} = \frac{1+k}{2k}\,\frac{\lambda^2}{\mu(\mu-\lambda)},$$

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

$$W = W_q + \frac{1}{\mu},$$

$$L = \lambda W.$$

Com vários atendentes ($M/E_k/s$), a relação da distribuição de Erlang para a distribuição exponencial que acabamos de descrever pode ser explorada para formular um processo de nascimento-e-morte *modificado* (cadeia de Markov de tempo contínuo) em termos de fases de atendimento exponenciais individuais ($k$ por cliente) em vez de clientes completos. Entretanto, não foi possível obter uma solução de estado estável genérica [quando $\rho = \lambda/(s\mu) < 1$] para a distribuição probabilística do número de clientes no sistema, conforme fizemos na Seção 17.5. Em vez disso, são necessárias teorias avançadas para resolver




■ FIGURA 17.9
Uma família de distribuições de Erlang com média constante $1/\mu$.

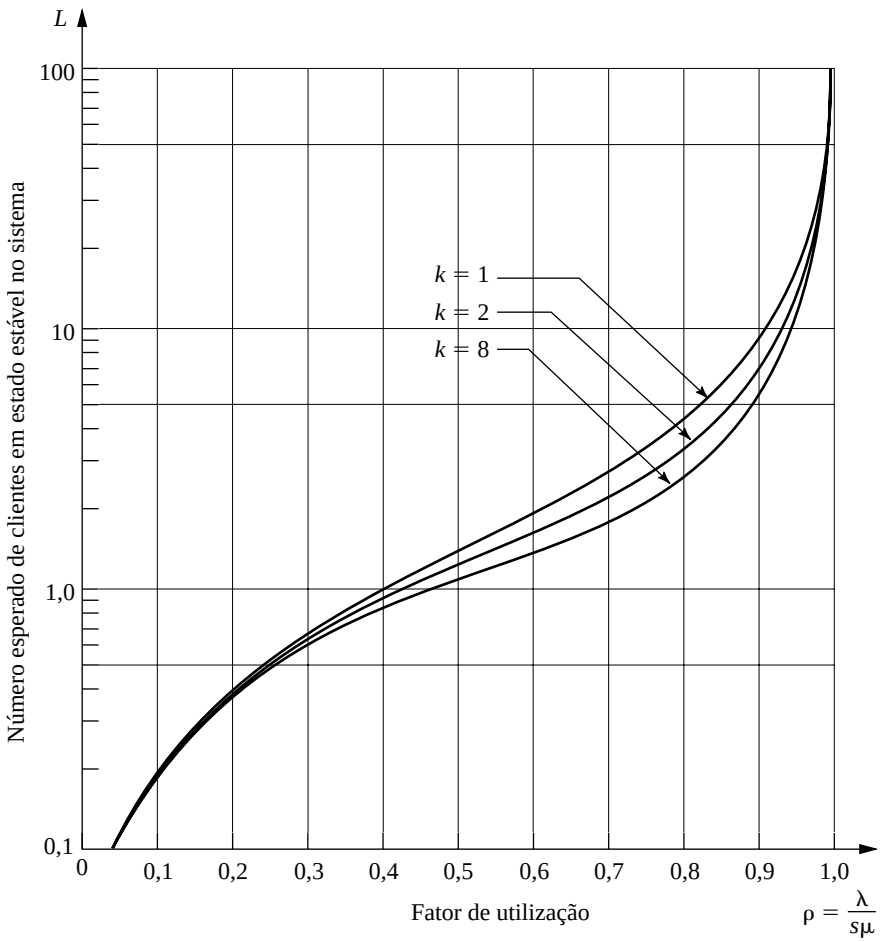
numericamente casos individuais. Repetindo, esses resultados foram obtidos e tabulados para inúmeros casos.[17] As médias ($L$) também são dadas graficamente na Figura 17.10 para alguns casos em que $s = 2$.

A seção de Exemplos Trabalhados do CD-ROM inclui um exemplo que aplica o modelo $M/E_k/s$ tanto para $s = 1$ quanto para $s = 2$ para escolher a alternativa de menor custo.

### Modelos sem uma Entrada de Poisson

Todos os modelos de filas apresentados até então partiram do pressuposto de que um processo de entrada de Poisson (tempos entre atendimentos exponenciais). Entretanto, essa hipótese é violada caso as chegadas sejam programadas ou reguladas de alguma maneira que as impeça de ocorrerem aleatoriamente, em cujo caso é necessário outro modelo.

Desde que os tempos de atendimento tenham uma distribuição exponencial com um parâmetro fixo, existem três desses modelos disponíveis. Esses modelos são obtidos meramente *invertendo-se* as supostas distribuições dos *tempos de atendimento* e *entre chegadas* nos três modelos precedentes. Portanto, o primeiro modelo novo ($GI/M/s$) não impõe nenhuma restrição sobre qual deva ser a distribuição de *tempos entre atendimentos*. Nesse caso, existem alguns resultados de estado estável disponíveis[18] (particularmente em relação às distribuições de tempos de espera) tanto para a versão do modelo com um único atendente quanto para aquela com vários atendentes, porém esses resultados não são nem de perto tão convenientes quanto as expressões simples dadas para o modelo $M/G/1$. O




■ FIGURA 17.10
Valores de $L$ para o modelo $M/E_k/2$ (Seção 17.7).

[17] Ibid.

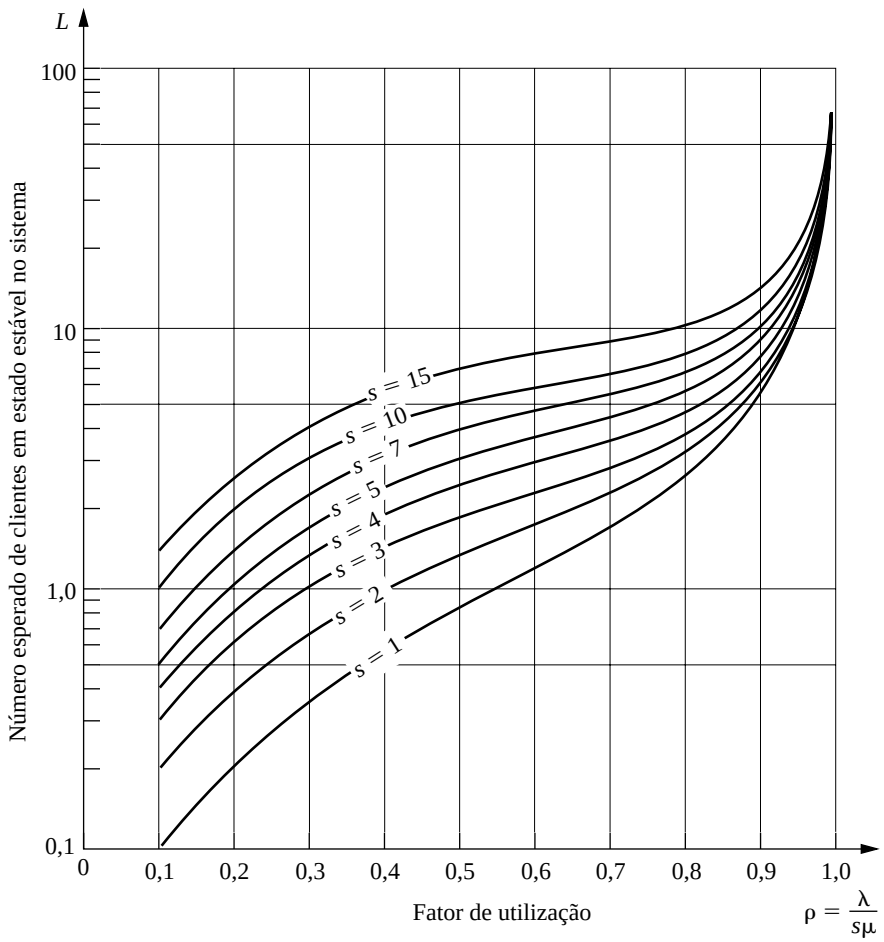
[18] Ver, por exemplo, as páginas 248-260 da Referência Selecionada 5.

segundo modelo novo ($D/M/s$) supõe que todos os tempos entre atendimentos sejam iguais a alguma *constante* fixa, que representaria um sistema de filas em que chegadas são *programadas* em intervalos regulares. O terceiro modelo novo ($E_k/M/s$) supõe uma distribuição de *Erlang* de tempos entre atendimentos, que fornece um meio-termo entre chegadas *regularmente programadas* (constantes) e *completamente aleatórias* (exponenciais). Foram tabulados[19] resultados computacionais extensos para esses dois últimos modelos, incluindo os valores de $L$ dados graficamente nas Figuras 17.11 e 17.12.

Se nem os tempos entre chegadas nem os tempos de atendimento para um sistema de filas tiverem uma distribuição exponencial, então existem outros três modelos de filas para os quais também estão disponíveis resultados computacionais.[20] Um desses modelos ($E_m/E_k/s$) supõe uma distribuição de Erlang para esses dois tipos de tempo. Os outros dois modelos ($E_k/D/s$ e $D/E_k/s$) supõem que um desses tempos tenham uma distribuição de Erlang e o outro tempo seja igual a alguma constante fixa.

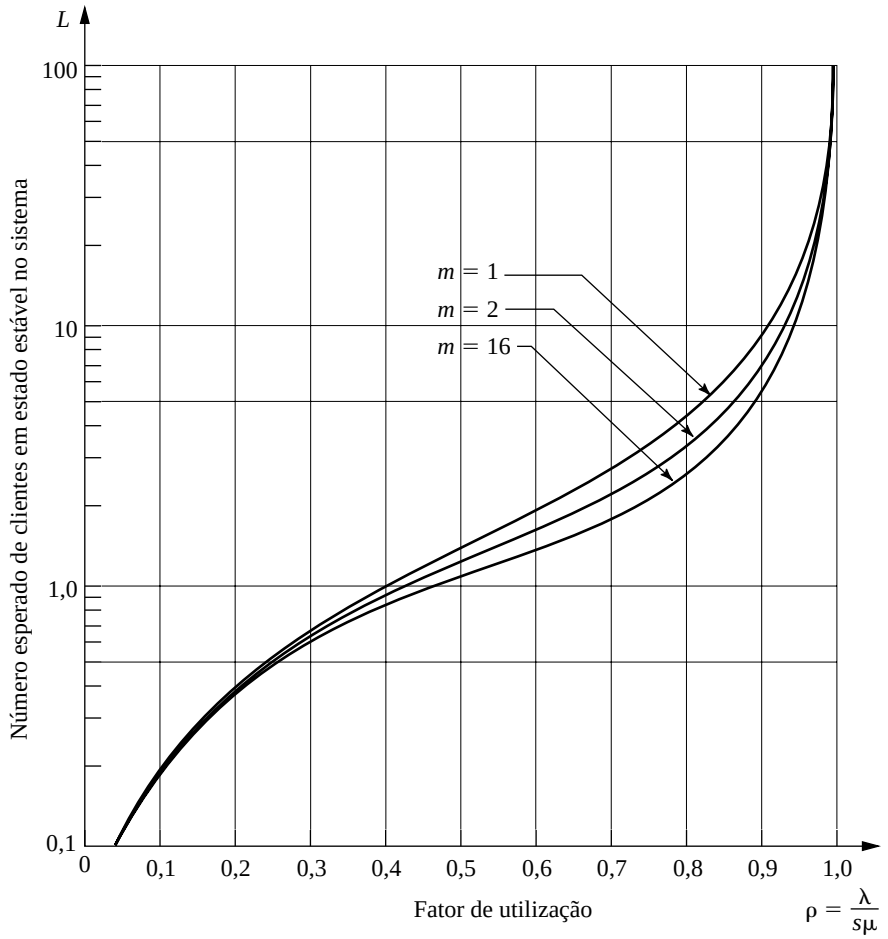
### Outros Modelos

Embora você tenha visto nesta seção um grande número de modelos de filas que envolvem distribuições não-exponenciais, estamos longe de ter esgotado a lista. Por exemplo, outra distribuição que é usada ocasionalmente, tanto para tempos entre atendimentos quanto para tempos entre chegadas, é a **distribuição hiperexponencial**. A característica-chave dessa




■ FIGURA 17.11
Valores de $L$ para o modelo $D/M/s$ (Seção 17.7).

[19] HILLIER; YU, op. cit.
[20] Ibid.




■ FIGURA 17.12
Valores de $L$ para o modelo $E_k/M/2$ (Seção 17.7).

distribuição é que, embora sejam permitidos apenas valores não-negativos, seu desvio-padrão $\sigma$, na verdade, é maior que sua média $1/\mu$. Essa característica contrasta com a distribuição de Erlang, em que $\sigma < 1/\mu$ em todos os casos, exceto para $k = 1$ (distribuição exponencial), que possui $\sigma = 1/\mu$. Para ilustrar uma situação típica na qual $\sigma > 1/\mu$ pode ocorrer, suponhamos que o atendimento envolvido no sistema de filas seja o reparo de algum tipo de máquina ou veículo. Se por acaso muitos reparos se tornarem uma rotina (tempos de atendimento pequenos), porém reparos ocasionais exigirem uma revisão geral (tempos de atendimento muito grandes), então o desvio-padrão dos tempos de atendimento tenderá a ser bastante grande em relação à média, em cujo caso a distribuição hiperexponencial pode ser usada para representar a distribuição de tempos de atendimento. Especificamente, essa distribuição suporia que existam probabilidades fixas, $p$ e $(1 - p)$, cujo tipo de reparo vai ocorrer, que o tempo necessário para cada tipo tem uma distribuição exponencial, porém que os parâmetros para essas duas distribuições exponenciais são diferentes. (Em geral, a distribuição hiperexponencial é tal qual um composto de duas ou mais distribuições exponenciais.)

Outra família de distribuições que está se tornando popular é aquela das **distribuições do tipo-fase** (algumas das quais também são chamadas *distribuições erlangianas generalizadas*). Essas distribuições são obtidas subdividindo-se o tempo total em um número de fases, cada uma tendo uma distribuição exponencial, na qual os parâmetros dessas distribuições exponenciais podem ser diferentes e as fases poderiam ser em série ou em paralelo (ou então ambas). Um grupo de fases *em paralelo* significa que o processo seleciona aleatoriamente *uma* das fases para percorrer de cada vez de acordo com probabilidades especificadas. Essa abordagem é, na realidade, como a distribuição hiperexponencial é

obtida, de modo que essa distribuição seja um caso especial das distribuições tipo-fase. Outro caso especial é a distribuição de Erlang, que tem as restrições de que todas suas *k* fases estão em série e que essas fases têm o *mesmo* parâmetro para suas distribuições exponenciais. Eliminar essas restrições significa que as distribuições tipo-fase, em geral, são capazes de fornecer consideravelmente maior flexibilidade que a distribuição de Erlang em adequar a verdadeira distribuição de tempos entre atendimentos ou de tempos de atendimento observada em um sistema de filas real. Essa flexibilidade é especialmente valiosa quando usar a distribuição real diretamente no modelo não for analiticamente tratável e a razão entre *média* e *desvio-padrão* para a distribuição real não se aproximar muito das razões disponíveis ($\sqrt{k}$ para $k = 1, 2, \ldots$) para a distribuição de Erlang.

Já que elas são construídas de combinações de distribuições exponenciais, os modelos de filas que usam distribuições tipo-fase ainda podem ser representados por uma *cadeia de Markov de tempo contínuo.* Essa cadeia de Markov geralmente terá um número de estados infinito, de forma que encontrar a distribuição de estado estável do estado do sistema requer resolver um sistema de equações lineares infinito com uma estrutura relativamente complexa. Resolver um sistema destes está longe de ser uma coisa rotineira, porém avanços teóricos recentes nos permitiram solucionar esses modelos de filas numericamente em alguns casos. Uma extensa tabulação desses resultados para modelos com várias distribuições tipo-fase (inclusive a distribuição hiperexponencial) está disponível.[21]

## 17.8 MODELOS DE FILAS DE DISCIPLINA DE PRIORIDADES

Em modelos de filas de disciplina de prioridades, a disciplina da fila se baseia em um *sistema de prioridades.* Portanto, a ordem na qual os membros da fila são selecionados se baseia nas prioridades que lhes foram atribuídas.

Muitos sistemas de filas reais se encaixam nesses modelos de disciplina de prioridades de forma muito mais aproximada do que para outros modelos disponíveis. Tarefas urgentes são colocadas à frente de outras tarefas e clientes importantes podem ter prioridade em relação a outros. Assim, o uso de modelos de disciplina de prioridade normalmente fornece um refinamento adequado em relação a outros modelos de filas mais usuais.

Apresentamos dois modelos básicos de disciplina de prioridades. Já que ambos os modelos fazem as mesmas hipóteses, exceto pela natureza das prioridades, descreveremos os modelos em conjunto e depois resumiremos seus resultados separadamente.

### Modelos

Ambos os modelos supõem que existam *N classes de prioridade* (a classe 1 tem a prioridade mais alta e a classe *N*, a mais baixa) e que sempre que um atendente ficar livre para começar a atender um novo cliente da fila, o cliente selecionado será aquele membro de classe de prioridade *mais alta* representada na fila por aquele que está esperando há mais tempo. Em outras palavras, os clientes são selecionados para começar a ser atendidos na ordem de suas classes de prioridade, mas também em uma ordem na qual os primeiros que chegam serão os primeiros a ser atendidos dentro de cada classe de prioridade. Parte-se do pressuposto da existência de um *processo de entrada de Poisson* e tempos de atendimento *exponenciais* para cada classe de prioridades. Exceto para o caso especial considerado mais à frente, os modelos também fazem a hipótese um tanto restritiva de que o tempo de atendimento esperado seja o *mesmo* para todas as classes de prioridades. Entretanto, os modelos permitem efetivamente que a taxa média de chegada seja diferente entre as diversas classes de prioridades.

A distinção entre os dois modelos é se as prioridades são *não-preemptivas* ou *preemptivas.* Com **prioridades não-preemptivas**, um cliente que está sendo atendido não pode ser jogado de volta para a fila (preterido) se um cliente com prioridade maior entrar no sistema de filas.

[21] SEELEN, L. P. et al. *Tables for Multi-Server Queues.* Amsterdã: North-Holland, 1985.

Portanto, assim que um atendente tiver começado a atender um cliente, o atendimento tem de ser completado sem interrupção. O primeiro modelo supõe prioridades não-preemptivas.

Com **prioridades preemptivas**, o cliente de menor prioridade que está sendo atendido é *preterido* (jogado de volta para a fila) toda vez que um cliente com prioridade maior entrar no sistema de filas. Um atendente é, portanto, liberado para começar a atender imediatamente a nova chegada. Quando um atendente não consegue *terminar* um atendimento, o próximo cliente a começar a receber atendimento é selecionado exatamente como descrito no início desta subseção, de modo que um cliente preterido normalmente voltará a ser atendido novamente e, após um número suficiente de tentativas, finalmente acabará de ser atendido. Em virtude da propriedade de falta de memória da distribuição exponencial (ver Seção 17.4), não precisamos nos preocupar em definir o ponto no qual o atendimento começa quando um cliente preterido voltar a ser atendido; a distribuição do tempo de atendimento *restante sempre* é a mesma. Para qualquer outra distribuição de tempo de atendimento, é importante distinguir entre sistemas *preemptivos-retomados*, em que o atendimento para um cliente preterido é retomado no ponto onde foi interrompido, e sistemas *preemptivos-repetidos*, nos quais o atendimento tem de começar do início novamente. O segundo modelo supõe prioridades preemptivas.

Para ambos os modelos, se a distinção entre clientes em diferentes classes de prioridades for ignorada, a Propriedade 6 para a distribuição exponencial (ver Seção 17.4) implica que *todos* os clientes chegariam de acordo com um processo de entrada de Poisson. Além disso, todos os clientes têm a *mesma* distribuição exponencial para tempos de atendimento. Conseqüentemente, os dois modelos, na verdade, são idênticos ao modelo *M/M/s* estudado na Seção 17.6, *exceto* pela ordem na qual os clientes são atendidos. Portanto, quando contamos apenas o *número total* de clientes no sistema, a distribuição de estado estável para o modelo *M/M/s* também se aplica a ambos os modelos. Portanto, as fórmulas para $L$ e $L_q$ também são transferidas, assim como os resultados esperados de tempo de espera (pela fórmula de Little) $W$ e $W_q$, para um cliente selecionado aleatoriamente. O que muda é a *distribuição* dos tempos de espera, que foi obtida na Seção 17.6 sob a hipótese de uma disciplina de fila em que os primeiros que chegam serão os primeiros a ser atendidos. Com uma disciplina de prioridades, essa distribuição tem uma *variância* muito maior, pois os tempos de espera de clientes nas classes de prioridades mais altas tendem a ser muito menores daqueles regidos pela regra na qual os primeiros que chegam serão os primeiros a ser atendidos, ao passo que os tempos de espera nas classes de prioridades mais baixas tendem a ser muito maiores. Pelo mesmo motivo, a subdivisão do número total de clientes no sistema tende a ser desproporcionalmente tendenciosa para as classes de prioridades mais baixas. Porém, essa condição é apenas a razão para impor prioridades no sistema de filas em primeiro lugar. Queremos *melhorar* as *medidas de desempenho* para cada uma das classes de prioridades mais altas à custa de desempenho para as classes de prioridades mais baixas. Para determinar o nível de melhoria que está sendo alcançado, precisamos obter tais medidas na forma de *tempo de espera previsto no sistema* e *número de clientes esperado no sistema* para cada uma das classes de prioridades. Expressões para essas medidas são dadas a seguir para os dois modelos, um de cada vez.

### Resultados para o Modelo de Prioridades Não-preemptivas

Façamos que $W_k$ seja o tempo de espera previsto no estado estável no sistema (incluindo o tempo de atendimento) para um membro da classe de prioridades $k$. Então

$$W_k = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu}, \qquad \text{para } k = 1, 2, \ldots, N,$$

em que
$$A = s!\frac{s\mu - \lambda}{r^s}\sum_{j=0}^{s-1}\frac{r^j}{j!} + s\mu,$$

$$B_0 = 1,$$

$$B_k = 1 - \frac{\sum_{i=1}^{k}\lambda_i}{s\mu},$$

$s$ = número de atendentes,

$\mu$ = taxa média de atendimento por atendente ocupado,

$\lambda_i$ = taxa média de chegada for classe de prioridades $i$,

$$\lambda = \sum_{i=1}^{N} \lambda_i,$$

$$r = \frac{\lambda}{\mu}.$$

Esse resultado parte do pressuposto de que

$$\sum_{i=1}^{k} \lambda_i < s\mu,$$

de modo que a classe de prioridades $k$ possa alcançar uma condição de estado estável. A *fórmula de Little* ainda se aplica a classes de prioridades individuais, de forma que $L_k$, o número esperado no estado estável de membros da classe de prioridades $k$ no sistema de filas (inclusive aqueles que estão sendo atendidos), seja

$$L_k = \lambda_k W_k, \text{ para } k = 1, 2, \ldots, N.$$

Para determinar o tempo de espera previsto na fila (excluindo o tempo de atendimento) para a classe de prioridades $k$, simplesmente subtraia $1/\mu$ de $W_k$; o comprimento esperado da fila correspondente é novamente obtido multiplicando por $\lambda_k$. Para o caso especial em que $s = 1$, a expressão para $A$ reduz-se a $A = \mu^2/\lambda$.

No *Courseware* de PO, você poderá encontrar um gabarito em Excel para realizar os cálculos anteriores.

A seção de Exemplos Trabalhados do CD-ROM fornece um exemplo que ilustra a aplicação do modelo das prioridades não-preemptivas para determinar quantos tornos-revólver uma fábrica deveria ter quando as tarefas caem nas três classes de prioridades.

### Variante com um Único Atendente do Modelo de Prioridades Não-preemptivas

A hipótese dada anteriormente que o tempo de atendimento esperado $1/\mu$ é o mesmo para todas as classes de prioridades é bastante restritiva. Na prática, essa hipótese algumas vezes é violada em decorrência das diferenças nas exigências de atendimento para as diferentes classes de prioridades.

Felizmente, para o caso especial de um único atendente, é possível permitir tempos de atendimento esperado diferentes e ainda obter resultados úteis. Façamos que $1/\mu_k$ represente a média da distribuição exponencial de tempos de atendimento para a classe de prioridades $k$, de modo que

$\mu_k$ = taxa média de atendimento para a classe de prioridades $k$, para $k = 1, 2, \ldots, N$.

Depois, o tempo de espera previsto no estado estável no sistema para um membro de classe de prioridades $k$ é

$$W_k = \frac{a_k}{b_{k-1}b_k} + \frac{1}{\mu_k}, \qquad \text{para } k = 1, 2, \ldots, N,$$

em que $$a_k = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i^2},$$

$$b_0 = 1,$$

$$b_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i}$$

Esse resultado vale desde que

$$\sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} < 1,$$

o que permite que classes de prioridades $k$ atinjam uma condição de estado estável. A fórmula de Little pode ser usada conforme descrito antes para obter outras medidas de desempenho importantes para cada classe de prioridades.

### Resultados para o Modelo de Prioridades Preemptivas

Para o modelo de prioridades preemptivas, precisamos restaurar a hipótese que o tempo de atendimento esperado é o mesmo para todas as classes de prioridades. Usando a mesma notação daquela utilizada no modelo original de prioridades não-preemptivas, fazendo que a preempção mude o tempo de espera *total* previsto no sistema (incluindo o tempo total de atendimento) para

$$W_k = \frac{1/\mu}{B_{k-1}B_k}, \qquad \text{para } k = 1, 2, \ldots, N,$$

para o caso *com um único atendente* ($s = 1$). Quando $s > 1$, $W_k$ pode ser calculado por um procedimento iterativo que será ilustrado no exemplo do Hospital Municipal. Os $L_k$ continuam a satisfazer a relação

$$L_k = \lambda_k W_k, \qquad \text{para } k = 1, 2, \ldots, N.$$

Os resultados correspondentes para a fila (excluindo os clientes que estão sendo atendidos) também podem ser obtidos de $W_k$ e $L_k$ exatamente como descrito para o caso das prioridades não-preemptivas. Em virtudes da propriedade da falta de memória da distribuição exponencial (ver Seção 17.4), as preempções não afetam o processo de atendimento (ocorrência de términos de atendimento) de qualquer maneira. O tempo de atendimento total previsto para qualquer cliente ainda é $1/\mu$.

O arquivo Excel deste capítulo inclui um gabarito em Excel para calcular as medidas de desempenho anterior para o caso com um único atendente.

### Exemplo do Hospital Municipal com Prioridades

No problema da sala de emergências do Hospital Municipal, o administrador percebeu que os pacientes não são tratados segundo a regra dos primeiros que chegam serão os primeiros a ser atendidos. Em vez disso, a enfermeira que recepciona os pacientes que chegam os divide, basicamente, em três categorias: (1) casos *críticos*, nos quais o pronto atendimento é vital para a sobrevivência do paciente; (2) casos *graves*, cujo tratamento prévio é importante para impedir maior agravamento; e (3) casos *estáveis*, em que o tratamento pode ser retardado sem conseqüências médicas adversas. Os pacientes são então tratados nessa ordem de prioridade, em que aqueles na mesma categoria são normalmente admitidos de acordo com a regra dos primeiros que chegam serão os primeiros a ser atendidos. Um médico interromperá o tratamento de um paciente caso surja um novo caso em uma categoria de maior prioridade. Aproximadamente 10% dos pacientes recaem na primeira categoria, 30% na segunda e 60% na terceira. Como os casos mais graves serão enviados ao hospital para cuidados posteriores após receber tratamento de emergência, o tempo de tratamento médio gasto por um médico na sala de emergências na verdade não difere muito entre essas categorias.

O administrador decidiu usar um modelo de filas de disciplina de prioridades como uma representação razoável desse sistema de filas, em que as três categorias de pacientes constituem as três classes de prioridades no modelo. Como o tratamento é interrompido pela chegada de um caso de prioridade mais alta, o *modelo de prioridades preemptivas* é o indicado. Tendo em vista os dados previamente disponíveis ($\mu = 3$ e $\lambda = 2$), as porcentagens

anteriores resultam em $\lambda_1 = 0{,}2$, $\lambda_2 = 0{,}6$ e $\lambda_3 = 1{,}2$. A Tabela 17.3 fornece os tempos de espera previstos na fila resultantes (e, assim, *excluindo* o tempo de tratamento) para as respectivas classes de prioridades[22] quando há um ($s = 1$) ou dois ($s = 2$) médicos de plantão. Os resultados correspondentes para o modelo de prioridades não-preemptivas também são dados na Tabela 17.3 para mostrar o efeito de preempção.

**Obtendo os Resultados da Prioridade Preemptiva.** Esses resultados de prioridade preemptiva para $s = 2$ foram obtidos como se segue. Como os tempos de espera para clientes da classe de prioridade 1 não são de modo algum afetados pela presença de clientes nas classes de prioridades menores, $W_1$ será o mesmo para quaisquer outros valores de $\lambda_2$ e $\lambda_3$, inclusive $\lambda_2 = 0$ e $\lambda_3 = 0$. Portanto, $W_1$ tem de ser igual a $W$ para o modelo com apenas *uma classe* correspondente (o modelo *M/M/s* na Seção 17.6) com $s = 2$, $\mu = 3$ e $\lambda = \lambda_1 = 0{,}2$, que resulta em

$$W_1 = W = 0{,}33370 \text{ hora}, \qquad \text{para } \lambda = 0{,}2$$

portanto

$$W_1 - \frac{1}{\mu} = 0{,}33370 - 0{,}33333 = 0{,}00037 \text{ hora}.$$

Consideremos agora as duas primeiras classes de prioridades. Observe novamente que os clientes nessas classes não são de forma alguma afetados pelas classes de prioridades mais baixas (somente classe de prioridade 3 nesse caso), que podem, conseqüentemente, ser ignorados na análise. Façamos que $\overline{W}_{1-2}$ seja o tempo de espera previsto no sistema (e, portanto, incluindo tempo de atendimento) de uma *chegada aleatória* em *qualquer* uma dessas duas classes, de modo que a probabilidade seja $\lambda_1/(\lambda_1 + \lambda_2) = \frac{1}{4}$ de que essa chegada se encontre na classe 1 e $\lambda_2/(\lambda_1 + \lambda_2) = \frac{3}{4}$ de que ela se encontre na classe 2. Portanto,

$$\overline{W}_{1-2} = \frac{1}{4}W_1 + \frac{3}{4}W_2.$$

Além disso, como o tempo de espera *previsto* é o mesmo para *qualquer* disciplina da fila, $\overline{W}_{1-2}$ também deve ser igual a $W$ para o modelo *M/M/s* na Seção 17.6, com $s = 2$, $\mu = 3$ e $\lambda = \lambda_1 + \lambda_2 = 0{,}8$, que resulta em

TABELA 17.3 **Resultados de estado estável dos modelos de disciplina de prioridades para o problema do Hospital Municipal**

| | Prioridades Preemptivas | | Prioridades Não-preemptivas | |
|---|---|---|---|---|
| | $s = 1$ | $s = 2$ | $s = 1$ | $s = 2$ |
| $A$ | — | — | 4,5 | 36 |
| $B_1$ | 0,933 | — | 0,933 | 0,967 |
| $B_2$ | 0,733 | — | 0,733 | 0,867 |
| $B_3$ | 0,333 | — | 0,333 | 0,667 |
| $W_1 - \frac{1}{\mu}$ | 0,024 hora | 0,00037 hora | 0,238 hora | 0,029 hora |
| $W_2 - \frac{1}{\mu}$ | 0,154 hora | 0,00793 hora | 0,325 hora | 0,033 hora |
| $W_3 - \frac{1}{\mu}$ | 1,033 hora | 0,06542 hora | 0,889 hora | 0,048 hora |

[22] Note que esses tempos esperados não podem mais ser interpretados como o tempo esperado antes de o tratamento começar quando $k < 1$, pois o tratamento poderia ser interrompido pelo menos uma vez, provocando tempo de espera adicional antes de o atendimento ser completado.

$\overline{W}_{1-2} = W = 0{,}33937$ hora, para $\lambda = 0{,}8$.

Combinando esses dois fatos resulta em

$$W_2 = \frac{4}{3}\left[0{,}33937 - \frac{1}{4}(0{,}33370)\right] = 0{,}34126 \text{ hora.}$$

$$\left(W_2 - \frac{1}{\mu} = 0{,}00793 \text{ hora.}\right)$$

Finalmente, façamos que $\overline{W}_{1-3}$ seja o tempo de espera previsto no sistema (e, assim, incluindo o tempo de atendimento) para uma *chegada aleatória* em *qualquer* uma das três classes de prioridades, de modo que as probabilidades são 0,1, 0,3 e 0,6 que se encontram, respectivamente, nas classes 1, 2 e 3. Portanto,

$$\overline{W}_{1-3} = 0{,}1W_1 + 0{,}3W_2 + 0{,}6W_3.$$

Além disso, $\overline{W}_{1-3}$ também tem de ser igual a $W$ para o modelo *M/M/s* na Seção 17.6, com $s = 2$, $\mu = 3$ e $\lambda = \lambda_1 + \lambda_2 + \lambda_3 = 2$, de modo que (da Tabela 17.2)

$\overline{W}_{1-3} = W = 0{,}375$ hora, para $\lambda = 2$.

Conseqüentemente,

$$W_3 = \frac{1}{0.6}[0{,}375 - 0{,}1(0{,}33370) - 0{,}3(0{,}34126)]$$
$$= 0{,}39875 \text{ hora.}$$

$$\left(W_3 - \frac{1}{\mu} = 0{,}06542 \text{ hora.}\right)$$

Os resultados $W_q$ correspondentes para o modelo *M/M/s* na Seção 17.6 também poderiam ter sido usados exatamente da mesma forma para se obter diretamente os valores $W_k - 1/\mu$.

**Conclusões.** Quando $s = 1$, os valores $W_k - 1/\mu$ da Tabela 17.3 para o caso das prioridades preemptivas indicam que disponibilizar apenas um médico faria que casos críticos teriam de aguardar $1\frac{1}{2}$ minuto (0,024 hora) em média, casos graves precisariam esperar mais de 9 minutos e casos estáveis deveriam esperar mais de 1 hora. Compare esses resultados com a espera média de $W_q = \frac{2}{3}$ hora para todos os pacientes que foi obtida na Tabela 17.2 segundo a disciplina de fila na qual os primeiros que chegam serão os primeiros a ser atendidos. Entretanto, esses valores representam *expectativas estatísticas,* de modo que alguns pacientes terão de esperar consideravelmente mais que a média para suas classes de prioridades. Essa demora não seria tolerável para os casos críticos e graves, nos quais alguns poucos minutos podem ser vitais. Ao contrário, os resultados com $s = 2$ da Tabela 17.3 (caso das prioridades preemptivas) indicam que acrescentar um segundo médico praticamente eliminaria a espera para todos, exceto os casos estáveis. Portanto, o administrador recomendou a presença de dois médicos de plantão na sala de emergências durante as primeiras horas da noite no próximo ano. A diretoria do Hospital Municipal adotou essa recomendação e simultaneamente aumentou o ônus para uso da sala de emergências!

## 17.9 REDES DE FILAS

Até agora consideramos apenas sistemas de filas com uma *única* instalação de atendimento com um ou mais atendentes. Entretanto, sistemas de filas encontrados em estudos de PO são algumas vezes, na realidade, *redes de filas*, isto é, redes de instalações de atendimento onde clientes devem receber atendimento em algumas ou todas essas instalações. Por exemplo, pedidos que estão sendo processados em uma ferramentaria devem ser direcionados por

meio de uma seqüência de grupos de máquinas (instalações de atendimento). Assim, é necessário estudar toda a rede para obter informações como o tempo de espera previsto total, número de clientes esperados no sistema todo e assim por diante.

Em virtude da importância de redes de filas, as pesquisas nessa área estão muito ativas. Entretanto, esta é uma área difícil, de modo que nos limitaremos a uma breve introdução.

Um desses resultados é de tal importância para redes de filas que essa descoberta e suas implicações merecem especial atenção aqui. Esse resultado fundamental é a *propriedade de equivalência* a seguir para o *processo de entrada* de clientes que chegam e o *processo de saída* de clientes que saem para certos sistemas de filas.

> **Propriedade da equivalência:** Suponha que uma instalação de atendimento com $s$ atendentes e uma fila infinita tenha uma entrada de Poisson com parâmetro $\lambda$ e a mesma distribuição exponencial de tempo de atendimento com parâmetro $\mu$ para cada atendente (o modelo *M/M/s*), em que $s\mu > \lambda$. Então a *saída* de estado estável dessa instalação de atendimento também é um processo de Poisson[23] com parâmetro $\lambda$.

Note como essa propriedade não faz nenhuma hipótese em relação ao tipo de disciplina da fila usada. Seja ela os primeiros que chegam serão os primeiros a ser atendidos, aleatória ou até mesmo uma disciplina de prioridades como indicado na Seção 17.8, os clientes atendidos deixarão a instalação de atendimento de acordo com um processo de Poisson. A implicação crucial desse fato para redes de filas é que se esses clientes tiverem de ir a outra instalação de atendimento para outros atendimentos, essa segunda instalação *também* terá uma entrada de Poisson. Com uma distribuição exponencial de tempos de atendimento, a propriedade de equivalência também será válida para essa instalação, que poderá então fornecer uma entrada de Poisson para uma terceira instalação etc. Discutimos a seguir as conseqüências para dois tipos de redes.

### Filas Infinitas em Séries

Suponha que todos os clientes tenham de receber atendimento em uma *série* de $m$ instalações de atendimento em uma seqüência fixa. Suponha que cada instalação tenha uma fila infinita (nenhuma limitação no número de clientes permitidos na fila), de modo que a série de instalações forma um sistema de *filas infinitas em série.* Suponha ainda que os clientes cheguem na primeira instalação de acordo com um processo de Poisson com parâmetro $\lambda$ e que cada instalação $i$ ($i = 1, 2, \ldots, m$) tenha uma distribuição exponencial de tempos de atendimento com parâmetro $\mu_i$ para seus $s_i$ atendentes, em que $s_i\mu_i > \lambda$. Decorre então da propriedade da equivalência que (sob condições de estado estável) cada instalação de atendimento tem uma entrada de Poisson com parâmetro $\lambda$. Portanto, o *modelo M/M/s* elementar da Seção 17.6 (ou seus equivalentes de disciplina de prioridades da Seção 17.8) pode ser usado para analisar cada instalação de atendimento independentemente dos demais!

Ser capaz de usar o modelo *M/M/s* para obter todas as medidas de desempenho para cada instalação independentemente, em vez de analisar interações entre instalações, é uma simplificação tremenda. Por exemplo, a probabilidade de ter $n$ clientes em dada instalação é dada pela fórmula para $P_n$ na Seção 17.6 para o modelo *M/M/s*. A *probabilidade conjunta* de $n_1$ clientes na instalação 1, $n_2$ clientes na instalação 2, . . . , então, é o *produto* das probabilidades individuais obtidas nessa maneira simples. Particularmente, essa probabilidade conjunta pode ser expressa como

$$P\{(N_1, N_2, \ldots, N_m) = (n_1, n_2, \ldots, n_m)\} = P_{n1}P_{n2}\ldots P_{nm}.$$

Essa forma simples para a solução é chamada **solução em forma de produto**. De maneira similar, o tempo de espera previsto total e o número esperado de clientes no sistema inteiro podem ser obtidos meramente somando os valores correspondentes obtidos nas respectivas instalações.

---

[23] Para uma demonstração, ver BURKE, P. J. The Output of a Queueing System. *Operations Research,* v. 4, n. 6, p. 699-704, 1956.

Infelizmente, a propriedade da equivalência e suas implicações não são válidas para o caso de *filas finitas* discutido na Seção 17.6. Esse caso é, na verdade, bem importante na prática, pois muitas vezes existe uma limitação definida no comprimento da fila em frente das instalações de atendimento em redes. Por exemplo, somente uma pequena quantidade de espaço de armazenagem em *buffer* é fornecida tipicamente em frente de cada instalação (estação) em um sistema de linha de produção. Para sistemas de filas finitas em série desse tipo, não existe uma solução em forma de produto simples. As instalações devem, sim, ser analisadas em conjunto e até agora só foram obtidos resultados limitados.

### Redes de Jackson

Os sistemas de filas infinitas em série não são as únicas redes de filas em que o modelo *M/M/s* pode ser usado para analisar cada instalação de atendimento independentemente das demais. Outro tipo proeminente de rede com essa propriedade (uma solução em forma de produto) é a *rede de Jackson,* nome dado em homenagem ao responsável por ter caracterizado a rede pela primeira vez e ter demonstrado que essa propriedade era válida.[24]

As características de uma rede de Jackson são as mesmas supostas anteriormente para o sistema de filas infinitas em série, exceto que agora os clientes visitam as instalações em ordens diferentes (sendo possível que eles não visitem todas elas). Para cada instalação, seus clientes que chegam provêm *tanto* de fora do sistema (de acordo com um processo de Poisson) quanto de outras instalações. Essas características são sintetizadas a seguir.

Uma **rede de Jackson** é um sistema de $m$ instalações de atendimento onde a instalação $i$ ($i = 1, 2, \ldots, m$) tem

**1.** Uma fila infinita.
**2.** Clientes provenientes de fora do sistema de acordo com um processo de entrada de Poisson com parâmetro $a_i$.
**3.** $s_i$ atendentes com uma distribuição exponencial de tempos de atendimento com parâmetro $\mu_i$.

Um cliente deixando a instalação $i$ é direcionado para a instalação $j$ ($j = 1, 2, \ldots, m$) com probabilidade $p_{ij}$ ou deixa o sistema com probabilidade

$$q_i = 1 - \sum_{j=1}^{m} p_{ij}.$$

Qualquer rede desse tipo possui a seguinte propriedade fundamental.

Sob condições de estado estável, cada instalação $j$ ($j = 1, 2, \ldots, m$) em uma rede de Jackson se comporta como se fosse um sistema de filas *M*/*M*/*s independente* com taxa de chegada

$$\lambda_j = a_j + \sum_{i=1}^{m} \lambda_i p_{ij},$$

em que $s_j\mu_j > \lambda_j$.

Essa propriedade fundamental não pode ser *provada* diretamente da propriedade da equivalência desta vez (o raciocínio se tornaria circular), mas seu *respaldo intuitivo* ainda é fornecido pela última propriedade. O ponto de vista intuitivo (tecnicamente não muito correto) é que, para cada instalação $i$, seus processos de entrada das diversas fontes (externas e de outras instalações) são *processos de Poisson independentes,* de modo que o *processo de entrada agregado* seja de Poisson com parâmetro $\lambda_i$ (a Propriedade 6 da Seção 17.4). A propriedade da equivalência diz então que o processo de *saída agregado* para a instalação $i$ tem de ser de Poisson com parâmetro $\lambda_i$. Desagregando esse processo de saída (novamente Propriedade 6), o processo para clientes que vão da instalação $i$ para a instalação $j$ deve ser de Poisson com parâmetro $\lambda_i p_{ij}$. Esse processo se torna um dos processos de entrada de Poisson para a instalação $j$, ajudando portanto a manter a série de processos de Poisson no sistema como um todo.

[24] Ver JACKSON, J. R. Jobshop-Like Queueing Systems. *Management Science,* v. 10, n. 1, p. 131-142, 1963.

A equação dada para obter $\lambda_j$ se baseia no fato que $\lambda_i$ é a *taxa de partida*, bem como a taxa de chegada para todos os clientes usando a instalação *i*. Como $p_{ij}$ é a proporção de clientes partindo da instalação *i* que em seguida vão para a instalação *j*, a taxa na qual os clientes da instalação *i* chegam na instalação *j* é $\lambda_i p_{ij}$. Somar esse produto ao longo de todos os *i*, e depois acrescentando essa soma a $a_j$, fornece a *taxa de chegada total* para a instalação *j* proveniente de todas as fontes.

Calcular $\lambda_j$ a partir dessa equação requer saber os $\lambda_i$ para $i \neq j$, porém esses $\lambda_i$ também são desconhecidos dados pelas equações correspondentes. Portanto, o procedimento é encontrar *simultaneamente* $\lambda_1, \lambda_2, \ldots, \lambda_m$, obtendo a solução simultânea de todo o sistema de equações lineares para $\lambda_j$ para $j = 1, 2, \ldots, m$. O Tutorial IOR inclui um procedimento interativo para encontrar $\lambda_j$ dessa maneira.

Para ilustrar esses cálculos, considere uma rede de Jackson com três instalações de atendimento com parâmetros mostrados na Tabela 17.4. Agregando a fórmula para $\lambda_j$ para $j = 1, 2, 3$, obtemos

$$\begin{aligned}\lambda_1 &= 1 &&+ 0{,}1\lambda_2 &&+ 0{,}4\lambda_3\\ \lambda_2 &= 4 + 0{,}6\lambda_1 &&&&+ 0{,}4\lambda_3\\ \lambda_3 &= 3 + 0{,}3\lambda_1 &&+ 0{,}3\lambda_2.\end{aligned}$$

Raciocine com base em cada equação para ver por que ela fornece a taxa de chegada total para a instalação correspondente. A solução simultânea para esse sistema é

$$\lambda_1 = 5,\ \lambda_2 = 10,\ \lambda_3 = 7\frac{1}{2}.$$

Dada essa solução simultânea, cada uma das três instalações de atendimento agora pode ser analisada *independentemente* usando as fórmulas para o modelo *M/M/s* dado na Seção 17.6. Por exemplo, para obter a distribuição do número de clientes $N_i = n_i$ na instalação *i*, note que

$$\rho_i = \frac{\lambda_i}{s_i\mu_i} = \begin{cases}\dfrac{1}{2} & \text{para } i = 1\\[2mm] \dfrac{1}{2} & \text{para } i = 2\\[2mm] \dfrac{3}{4} & \text{para } i = 3.\end{cases}$$

Agregando esses valores (e os parâmetros da Tabela 17.4) na fórmula para $P_n$ resulta em

$$P_{n_1} = \frac{1}{2}\left(\frac{1}{2}\right)^{n_1} \qquad \text{para a instalação 1,}$$

$$P_{n_2} = \begin{cases}\dfrac{1}{3} & \text{para } n_2 = 0\\[2mm] \dfrac{1}{3} & \text{para } n_2 = 1\\[2mm] \dfrac{1}{3}\left(\dfrac{1}{2}\right)^{n_2-1} & \text{para } n_2 \geq 2\end{cases} \qquad \text{para a instalação 2,}$$

$$P_{n_3} = \frac{1}{4}\left(\frac{3}{4}\right)^{n_3} \qquad \text{para a instalação 3.}$$

TABELA 17.4 **Dados para o exemplo de uma rede de Jackson**

| Instalação $j$ | $s_j$ | $\mu_j$ | $a_j$ | $p_{ij}$ | | |
|---|---|---|---|---|---|---|
| | | | | $i = 1$ | $i = 2$ | $i = 3$ |
| $j = 1$ | 1 | 10 | 1 | 0 | 0,1 | 0,4 |
| $j = 2$ | 2 | 10 | 4 | 0,6 | 0 | 0,4 |
| $j = 3$ | 1 | 10 | 3 | 0,3 | 0,3 | 0 |

A *probabilidade conjunta* de $(n_1, n_2, n_3)$ é dada então simplesmente pela solução em forma de produto

$$P\{(N_1, N_2, N_3) = (n_1, n_2, n_3)\} = P_{n1}P_{n2}P_{n3}.$$

De forma similar, o número esperado de clientes $L_i$ na instalação $i$ pode ser calculado da Seção 17.6 como

$$L_1 = 1, \qquad L_2 = \frac{4}{3}, \qquad L_3 = 3.$$

O *número* esperado *total* de clientes no sistema todo é então

$$L = L_1 + L_2 + L_3 = 5\frac{1}{3}.$$

Obtendo $W$, o tempo de espera *total* previsto no sistema (incluindo tempos de atendimento) para um cliente, é um pouco mais capcioso. Não podemos simplesmente adicionar os tempos de espera previstos nas respectivas instalações, pois um cliente não visita necessariamente cada instalação exatamente uma vez. Entretanto, a fórmula de Little ainda pode ser usada, em que a taxa de chegada $\lambda$ ao sistema é a soma das taxas de chegada *provenientes de fora* das instalações, $\lambda = a_1 + a_2 + a_3 = 8$. Portanto,

$$W = \frac{L}{a_1 + a_2 + a_3} = \frac{2}{3}.$$

Concluindo, devemos indicar que realmente existem outros tipos (mais complicados) de redes de filas nas quais as instalações de atendimento individuais podem ser analisadas independentemente das demais. Na realidade, encontrar redes de filas com uma solução em forma de produto tem sido a sonhada meta da pesquisa sobre redes de filas. Algumas fontes adicionais de informação são as Referências Selecionadas 1, 2, 12 e 14.

## 17.10 APLICAÇÃO DA TEORIA DAS FILAS

Em razão da abundância de informações fornecida pela teoria das filas, ela é amplamente usada para orientar no projeto (ou redesenho) de sistemas de filas. Agora, podemos mudar nosso foco para como a teoria das filas é aplicada dessa forma.

A decisão mais comum que precisa ser feita ao desenhar um sistema de filas é quantos atendentes deverão ser disponibilizados. Entretanto, uma série de outras decisões também é necessária. Entre as possíveis decisões, temos

1. Número de atendentes em uma instalação de atendimento.
2. Eficiência dos atendentes.
3. Número de instalações de atendimento.
4. Dimensionamento do tempo de espera na fila.
5. Quaisquer prioridades para categorias de clientes diversas.

As duas considerações primárias na tomada dessas decisões são, tipicamente: (1) o custo da capacidade de atendimento fornecida pelo sistema de filas e (2) as conseqüências de se fazer os clientes esperarem no sistema de filas. Disponibilizar muita capacidade de atendimento provoca custos excessivos. Disponibilizar pouca capacidade provoca espera excessiva. Portanto, a meta é encontrar um equilíbrio entre custo de atendimento e tempo de espera.

Existem duas abordagens básicas para procurar alcançar esse equilíbrio. Uma é estabelecer um ou mais critérios para um nível de atendimento satisfatório em termos de quanto tempo de espera seria aceitável. Por exemplo, um critério possível poderia ser que o tempo de espera previsto no sistema não poderia exceder determinado número de minu-

tos. Outro poderia ser que pelo menos 95% dos clientes deveriam esperar não mais de certo número de minutos no sistema. Critérios similares em termos do número de clientes previstos no sistema (ou a distribuição probabilística desse número) também poderiam ser usados. Os critérios também poderiam ser colocados em termos do tempo de espera ou do número de clientes na *fila* em vez de no sistema. Assim que o critério ou critérios tiverem sido selecionados, então normalmente é simples usar um método de tentativa e erro para encontrar o desenho do sistema de filas menos oneroso que satisfaça todos os critérios.

A outra abordagem básica para procurar o melhor equilíbrio envolve avaliar os custos associados às conseqüências de se fazer os clientes esperarem. Suponha, por exemplo, que o sistema de filas seja um *sistema de atendimento interno* (conforme descrito na Seção 17.3), no qual os clientes são os empregados de uma empresa que visa lucros. Fazer esses empregados esperarem no sistema de filas provoca *perda de produtividade,* o que resulta em *perda de lucros.* Essa perda de lucros é o **custo de espera** associado ao sistema de filas. Expressando esse custo de espera em função do tempo de espera, o problema de determinar o melhor desenho do sistema de filas agora pode ser colocado como minimizar o *custo total* esperado (custo de atendimento mais custo de espera) por unidade de tempo.

A seguir, explicamos em detalhes essa última abordagem para o problema de determinação do número ótimo de atendentes a serem disponibilizados.

### Quantos Atendentes Devem Ser Disponibilizados?

Para formular a função objetivo quando a variável de decisão é o número de atendentes $s$, façamos que

$E(\text{TC})$ = custo total esperado por unidade de tempo,

$E(\text{SC})$ = custo de atendimento esperado por unidade de tempo,

$E(\text{WC})$ = custo de espera esperado por unidade de tempo.

Então o objetivo é escolher o número de atendentes de modo a

Minimizar $E(\text{TC}) = E(\text{SC}) + E(\text{WC})$.

Quando cada um dos custos de atendimento for o mesmo, o **custo de atendimento** será

$$E(\text{SC}) = C_s s,$$

em que $C_s$ é o custo marginal de um atendente por unidade de tempo. Para calcular WC para qualquer valor de $s$, note que $L = \lambda W$ fornece o tempo total de espera previsto no sistema de filas por unidade de tempo. Assim, quando o custo de espera for proporcional ao tempo de espera, esse custo pode ser expresso como

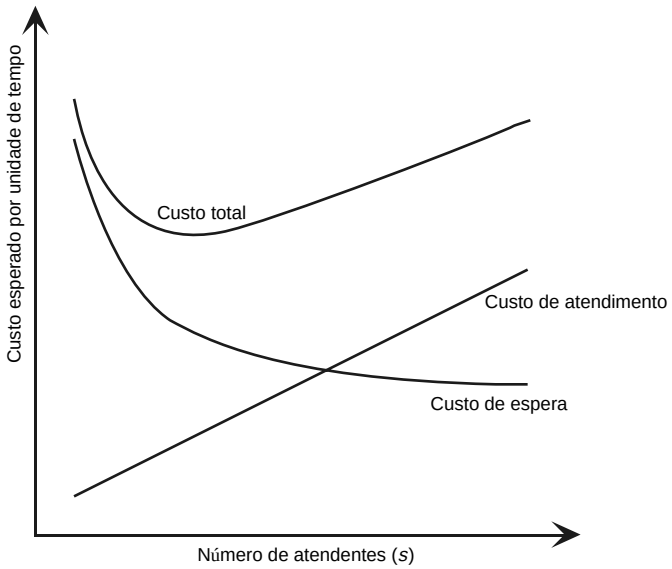
$$E(\text{WC}) = C_w L,$$

em que $C_w$ é o custo de espera por unidade de tempo para cada cliente no sistema de filas. Portanto, após estimar as constantes, $C_s$ e $C_w$, o objetivo é escolher o valor de $s$ de modo a

Minimizar $E(\text{TC}) = C_s s + C_w L$.

Escolhendo o modelo de filas que se ajuste ao sistema de filas, o valor de $L$ pode ser obtido para diversos valores de $s$. Aumentar $s$ diminui $L$, no início de forma rápida e depois gradualmente de forma mais lenta.

A Figura 17.13 mostra a forma geral das curvas $E(\text{SC})$, $E(\text{WC})$ e $E(\text{TC})$ *versus* o número de atendentes $s$. Para melhor conceitualização, desenhamos essas curvas como curvas suaves, embora os únicos valores viáveis de $s$ sejam $s = 1, 2, \ldots$. Calcular $E(\text{TC})$ para valores consecutivos de $s$ até que $E(\text{TC})$ pare de diminuir e, ao contrário, comece a aumentar é simples para encontrar o número de atendentes que minimize o custo total. O exemplo a seguir ilustra esse processo.




■ FIGURA 17.13
A forma das curvas de custos esperados para determinar o número de atendentes a serem disponibilizados.

### Exemplo

A Ferramentaria Acme tem um almoxarifado para armazenar ferramentas a serem usadas pelos ferramenteiros. Dois almoxarifes administram o almoxarifado. Esses almoxarifes distribuem as ferramentas à medida que os ferramenteiros chegam e as solicitam. Depois, essas ferramentas são devolvidas aos almoxarifes quando eles não precisarem mais delas. Têm havido reclamações dos supervisores de que seus ferramenteiros têm de perder muito tempo esperando ser atendidos no almoxarifado, de modo que parece ser necessário *maior* número de almoxarifes. No entanto, a gerência está exercendo pressão para reduzir os gastos indiretos na fábrica e essa redução levaria a um número *menor* de almoxarifes. Para solucionar essas pressões conflitantes, está sendo realizado um estudo de PO para determinar exatamente quantos almoxarifes deve ter o almoxarifado.

O almoxarifado forma um sistema de filas, no qual os almoxarifes são seus atendentes e os ferramenteiros seus clientes. Após coletar alguns dados sobre os tempos entre atendimentos e tempos de atendimento, a equipe de PO chegou à conclusão que o modelo de filas que melhor se ajusta a esse sistema de filas é o modelo *M/M/s*. As estimativas da taxa média de chegada $\lambda$ e da taxa média de atendimento (por atendente) $\mu$ são

$\lambda = 120$ clientes por hora,
$\mu = 80$ clientes por hora,

de modo que o fator de utilização para os dois almoxarifes seja

$$\rho = \frac{\lambda}{s\mu} = \frac{120}{2(80)} = 0{,}75.$$

O custo total para a companhia de cada almoxarife é cerca de US$ 20 por hora e, dessa forma, $C_s$ = US$ 20. Enquanto um ferramenteiro estiver ocupado, o valor para a companhia de suas produções médias é de US$ 48 por hora e, assim, $C_w$ = US$ 48. Portanto, a equipe de PO agora precisa encontrar o número de atendentes (almoxarifes) $s$ que vai

Minimizar $E(\text{TC}) = \text{US\$ } 20\, s + \text{US\$ } 48\, L$.

Existe um gabarito em Excel no *Courseware* de PO para calcular esses custos com o modelo *M/M/s*. Tudo o que é preciso fazer é introduzir os dados para o modelo juntamente com o custo de atendimento unitário $C_s$, o custo de espera unitário $C_w$ e o número de atendentes $s$ que você quiser tentar. O gabarito calcula então $E$(SC), $E$(WC) e $E$(TC). Isso está

ilustrado na Figura 17.14 com $s = 3$ para esse exemplo. Introduzindo repetidamente valores alternativos de $s$, o gabarito pode então revelar qual valor minimiza $E$(TC) em uma questão de segundos.

A Tabela 17.5 mostra os dados que seriam gerados desse gabarito repetindo esses cálculos para $s = 1, 2, 3, 4$ e 5. Já que o fator de utilização para $s = 1$ é $\rho = 1{,}5$, um único almoxarife seria incapaz de atender os clientes, de modo que essa opção seja descartada. Todos os valores de $s$ maiores são viáveis, porém $s = 3$ tem o menor custo total esperado. Além disso, $s = 3$ diminuiria o custo total esperado atual para $s = 2$ por US$ 61 por hora. Portanto, apesar da intenção atual da gerência em reduzir os gastos indiretos (que inclui o custo dos almoxarifes), a equipe de PO recomenda que um terceiro almoxarife seja colocado no almoxarifado. Note que essa recomendação diminuiria o fator de utilização para os almoxarifes de um já modesto 0,75 para 0,5. Entretanto, em virtude da grande melhoria na produtividade dos ferramenteiros (que são muito mais caros que os almoxarifes) pela diminuição de seus tempos de espera desperdiçados no almoxarifado, a gerência adota a recomendação.

■ FIGURA 17.14
Este gabarito em Excel para emprego de análise econômica para escolher o número de atendentes com o modelo *M/M/s* é aplicado aqui ao exemplo da Ferramentaria Acme com $s = 3$.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | **Análise Econômica do Exemplo da Ferramentaria Acme** | | | | | | |
| 2 | | | | | | | |
| 3 | | | **Dados** | | | | Resultados |
| 4 | | l = | 120 | (taxa média de chegada) | | L = | 1,736842105 |
| 5 | | m= | 80 | (taxa média de atendimento) | | $L_q$ = | 0,236842105 |
| 6 | | s = | s 3 | ( nº de atendentes) | | | |
| 7 | | | | | | W = | 0,014473684 |
| 8 | | Pr(𝒲> t) = | 0,02581732 | | | $W_q$ = | 0,001973684 |
| 9 | | quando t = | 0,05 | | | | |
| 10 | | | | | | $\rho$ = | 0,5 |
| 11 | | Prob(𝒲$_q$> t) = | 0,00058707 | | | | |
| 12 | | quando t = | 0,05 | | | n | $P_n$ |
| 13 | | | | | | 0 | 0,210526316 |
| 14 | | Análise Econômica: | | | | 1 | 0,315789474 |
| 15 | | $C_s$ = | US$ 20,00 | (custo / atendente / unidade de tempo) | | 2 | 0,236842105 |
| 16 | | $C_w$ = | US$ 48,00 | (custo de espera / unidade de tempo) | | 3 | 0,118421053 |
| 17 | | | | | | 4 | 0,059210526 |
| 18 | | Custo de Atendimento | US$ 60,00 | | | 5 | 0,029605263 |
| 19 | | Custo de Espera | US$ 83,37 | | | 6 | 0,014802632 |
| 20 | | Custo Total | US$ 143,37 | | | 7 | 0,007401316 |

| | B | C |
|---|---|---|
| 18 | Custo de Atendimento | =$C_s$*s |
| 19 | Custo de Espera | =$C_w$*L |
| 20 | Custo Total | =CustoDeAtendimento+CustoDeEspera |

| Nome da Faixa | Célula |
|---|---|
| CustoDeAtendimento | C18 |
| CustoDeEspera | C19 |
| $C_s$*s | C15 |
| $C_w$*L | C16 |
| **L** | **G4** |
| **s** | **C6** |
| CustoTotal | C20 |

TABELA 17.5 **Cálculo de *E*(TC) para alternativa s no exemplo da Ferramentaria Acme**

| $s$ | $\rho$ | $L$ | $E(SC) = C_s s$ | $E(WC) = C_w L$ | $E(TC) = E(SC) + E(WC)$ |
|---|---|---|---|---|---|
| 1 | 1,50 | ∞ | US$ 20 | ∞ | ∞ |
| 2 | 0,75 | 3,43 | US$ 40 | US$ 164,57 | US$ 204,57 |
| 3 | 0,50 | 1,74 | US$ 60 | US$ 83,37 | US$ 143,37 |
| 4 | 0,375 | 1,54 | US$ 80 | US$ 74,15 | US$ 154,15 |
| 5 | 0,30 | 1,51 | US$ 100 | US$ 72,41 | US$ 172,41 |

### Outras Questões

O Capítulo 26 do CD-ROM expande consideravelmente além da teoria das filas, inclusive como lidar com algumas outras questões não consideradas anteriormente.

Por exemplo, a análise da página 52 supunha que o custo de espera fosse proporcional ao tempo de espera, mas isso algumas vezes não é o caso. Se uma empresa tiver um ou dois de seus empregados em um sistema de filas, talvez isso não seja muito sério em termos da perda de produtividade deles, pois outros poderiam estar aptos a lidarem com todo o trabalho produtivo disponível. Entretanto, ter mais empregados no sistema de filas pode resultar em aumento agudo na perda de produtividade e o lucro perdido resultante, de modo que o custo de espera se torne uma função não-linear do número do sistema. Similarmente, as conseqüências para um sistema de atendimento comercial que fazem seus clientes esperarem podem ser mínimas para esperas curtas, porém muito mais graves para esperas longas. Nesse caso, o custo de espera se torna uma função não-linear do tempo de espera. A Seção 26.3 descreve a formulação de funções não-lineares de custo de espera e depois o cálculo de *E*(WC) com tais funções.

A Seção 26.4 discute um modelo de decisão em que as variáveis de decisão são *tanto* o número de atendentes quanto a taxa média de atendimento para os atendentes. Uma questão interessante que surge aqui é se é melhor ter *um atendente rápido* (várias pessoas trabalhando juntas para atender rapidamente cada cliente) ou *vários atendentes mais lentos* (várias pessoas trabalhando separadamente para atender clientes diferentes).

A Seção 26.4 também apresenta um modelo de decisão no qual as variáveis de decisão são o número de instalações de atendimento e o número de atendentes por instalação para fornecer atendimento para uma população solicitante de possíveis clientes. Dada a taxa média de chegada para toda a população solicitante, aumentar o número de instalações permite diminuir a média de chegada (carga de trabalho) a cada instalação. O número de instalações de atendimento também afeta quanto tempo cada cliente precisará despender na ida e na volta da instalação mais próxima. O custo de espera agora precisa ser uma função do tempo total perdido por um cliente esperando em uma instalação de atendimento ou indo e retornando da instalação. Portanto, a Seção 26.5 apresenta alguns modelos de tempo de viagem para determinar o tempo de viagem de ida e volta para cada cliente.

## 17.11 CONCLUSÕES

Sistemas de filas são dominantes na sociedade. A adequação desses sistemas pode ter um efeito importante sobre a qualidade de vida e produtividade.

A teoria das filas estuda sistemas de filas formulando modelos matemáticos de suas operações e depois como usar esses modelos para obter medidas de desempenho. Essa análise fornece informações vitais para desenhar, de forma eficaz, sistemas de filas que alcançam um equilíbrio apropriado entre o custo de fornecer um atendimento e o custo associado à espera por esse atendimento.

Este capítulo apresentou os modelos mais básicos da teoria das filas para os quais existem, particularmente, resultados úteis. Entretanto, muitos outros modelos interessantes

poderiam ser considerados caso o espaço permitisse. Na realidade, vários milhares de trabalhos de pesquisa formulando e/ou analisando modelos de filas já apareceram na literatura técnica e muitos mais estão sendo publicados a cada ano!

A *distribuição exponencial* desempenha papel fundamental na teoria das filas para representar a distribuição de tempos entre chegadas e tempos de atendimento, pois essa hipótese nos permite representar o sistema de filas como uma *cadeia de Markov de tempo contínuo.* Pela mesma razão, *distribuições tipo-fase* como a *distribuição de Erlang,* em que o tempo total é subdividido em fases individuais com uma distribuição exponencial, são muito úteis. Foram obtidos resultados analíticos úteis somente para um número relativamente pequeno de modelos de filas fazendo outras hipóteses.

Modelos de disciplina de prioridades de filas são úteis para a situação comum na qual algumas categorias de clientes recebem determinada prioridade em relação a outros que estão recebendo atendimento.

Em outra situação comum, clientes devem receber atendimento em diversas instalações de atendimento. Modelos para redes de filas estão ganhando ampla projeção para o emprego em tais situações. Essa é uma área de pesquisa em andamento especialmente ativa.

Quando não tivermos disponível nenhum modelo tratável que forneça uma representação razoável do sistema de filas em estudo, uma abordagem comum é obter dados de desempenho relevantes desenvolvendo um programa de computador para simular a operação do sistema. Essa técnica é discutida no Capítulo 20.

A Seção 17.10 descreve brevemente como a teoria das filas pode ser usada para ajudar a desenvolver sistemas de filas eficientes e depois o Capítulo 26 (no CD-ROM) expande consideravelmente esse tema.

## REFERÊNCIAS SELECIONADAS


**1.** CHAO, X. et al. *Queueing Networks:* Customers, Signals and Product Form. Nova York: Wiley, 1999.

**2.** CHEN, H.; YAO, D. D. *Fundamentals of Queueing Networks:* Performance, Asymptotics and Optimization. Nova York: Springer, 2001.

**3.** COOPER, R. B. *Introduction to Queueing Theory,* 2. ed. Nova York: Elsevier North-Holland, 1981. Também distribuído pelo Programa de Educação Continuada em Engenharia da George Washington University, Washington, DC.

**4.** _____. Queueing Theory. Capítulo 10. In: HEYMAN, D. P.; SOBLE, M. J. (Eds.). *Stochastic Models.* Amsterdã e Nova York: North Holland, 1990. Esse trabalho de pesquisa também é distribuído pelo Programa de Educação Continuada em Engenharia da George Washington University, Washington, DC.

**5.** GROSS, D.; HARRIS, C. M. *Fundamentals of Queueing Theory.* 3. ed. Nova York: Wiley, 1998.

**6.** HALL, R. W. *Queueing Methods:* For Services and Manufacturing. Upper Saddle River, NJ: Prentice-Hall, 1991.

**7.** HILLIER, F. S.; HILLIER, M. S. *Introduction to Management Science:* A Modeling and Case Studies Approach with Spreadsheets. 2. ed. Capítulo 14. Burr Ridge, IL: McGraw-Hill/Irwin, 2003.

**8.** KLEINROCK, L. *Queueing Systems. Vol. I:* Theory. Nova York: Wiley, 1975.

**9.** NORDGREN, B. The Problem with Waiting Times. *IIE Solutions,* p. 44-48, maio 1999.

**10.** PAPADOPOULOS, H. T. et al. *Queueing Theory in Manufacturing Systems Analysis and Design.* Londres: Chapman Hall, 1993.

**11.** PRABHU, N. U. *Foundations of Queueing Theory.* Boston: Kluwer Academic Publishers, 1997.

**12.** SERFOZO, R. *Introduction to Stochastic Networks.* Nova York: Springer, 1999.

**13.** STIDHAM JR., S. Analysis, Design and Control of Queueing Systems. *Operations Research,* v. 50, p. 197-216, 2002.

**14.** WALRAND, J. *An Introduction to Queueing Networks.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

**15.** WOLFF, R. W. *Stochastic Modeling and the Theory of Queues.* Englewood Cliffs, NJ: Prentice-Hall, 1989.

## FERRAMENTAS DE APRENDIZADO PARA ESTE CAPÍTULO INCLUÍDAS NO CD-ROM

**Exemplos Trabalhados:**

Exemplos para o Capítulo 17

**Procedimento Interativo no Tutorial IOR:**

Rede de Jackson

**Arquivos em Excel (Capítulo 17 — Teoria das Filas):**

Gabarito para o *Modelo M/M/s*
Gabarito para a Variante de Fila do Modelo *M/M/s*
Gabarito para a Variante da população solicitante Finita do Modelo *M/M/s*
Gabarito para o Modelo *M/G/*1
Gabarito para o Modelo *M/D/*1
Gabarito para o Modelo $M/E_k/1$
Gabarito para o Modelo de Prioridades Não-preemptivas
Gabarito para Modelo de Prioridades Preemptivas
Gabarito para a Análise de Econômica *M/M/s* do Número de Atendentes

**Arquivo Lingo (Capítulo 17 — Teoria das Filas) para Solucionar os Exemplos Selecionados**

**Glossário para o Capítulo 17**

Ver Apêndice 1 para obter documentação sobre o software.

## PROBLEMAS[25]

Inserimos um T à esquerda de alguns problemas (ou parte deles) toda vez que um dos gabaritos listados anteriormente puder ser útil. Um asterisco no número do problema indica que pelo menos há uma resposta parcial no final do livro.

**17.2-1.*** Considere uma barbearia típica. Demonstre que ela é um sistema de filas descrevendo seus componentes.

**17.2-2.*** João e José são dois barbeiros em uma barbearia que eles possuem e dirigem. Eles têm duas cadeiras de barbear para clientes que estão esperando por um corte de cabelo e, portanto, o número de clientes na barbearia varia entre 0 e 4. Para $n = 0, 1, 2, 3, 4$, a probabilidade $P_n$ de que exatamente $n$ clientes se encontrem na barbearia é $P_0 = \frac{1}{16}$, $P_1 = \frac{4}{16}$, $P_2 = \frac{6}{16}$, $P_3 = \frac{4}{16}$, $P_4 = \frac{1}{16}$.

**(a)** Calcule $L$. Como você descreveria o significado de $L$ para João e José?
**(b)** Para cada um dos possíveis valores do número de clientes no sistema de filas, especifique quantos clientes se encontram na fila. Calcule então $L_q$. Como você descreveria o significado de $L_q$ para João e José?
**(c)** Determine o número esperado de clientes atendidos.
**(d)** Dado que chegue uma média de quatro clientes por hora e permaneçam para cortar o cabelo, determine $W$ e $W_q$. Descreva esses dois valores em termos significativos para João e José.
**(e)** Dado que João e José são igualmente rápidos nos cortes de cabelo, qual é a duração média de um corte de cabelo?

**17.2-3.** A Mercearia Mom-and-Pop tem um pequeno estacionamento com três vagas reservadas para seus clientes. Durante o horário de funcionamento da mercearia, os carros entram no estacionamento e usam uma das vagas a uma taxa média de 2 por hora. Para $n = 0, 1, 2, 3$, a probabilidade $P_n$ de que exatamente $n$ vagas estejam sendo usadas no momento é $P_0 = 0{,}2$, $P_1 = 0{,}3$, $P_2 = 0{,}3$, $P_3 = 0{,}2$.

**(a)** Descreva como esse estacionamento pode ser interpretado como um sistema de filas. Particularmente, identifique os clientes e os atendentes. Qual é o atendimento que está sendo fornecido? O que constitui um tempo de atendimento? Qual é a capacidade da fila?
**(b)** Determine as medidas de desempenho básicas — $L$, $L_q$, $W$ e $W_q$ — para esse sistema de filas.
**(c)** Use os resultados do item (*b*) para determinar a duração média que um carro permanece no estacionamento.

**17.2-4.** Para cada uma das seguintes alternativas sobre a fila em um sistema de filas, classifique a alternativas como verdadeira ou falsa e, em seguida, justifique sua resposta referindo-se a uma afirmação específica no capítulo.

**(a)** Fila é onde clientes aguardam no sistema de filas até que seu atendimento seja completado..
**(b)** Modelos de filas supõem, convencionalmente, que a fila seja capaz de reter apenas um número limitado de clientes.
**(c)** A disciplina de fila mais comum é aquela na qual os primeiros que chegam serão os primeiros a ser atendidos.

[25] Consulte também o final do Capítulo 26 (no CD-ROM) para outros problemas envolvendo a aplicação da teoria das filas.

**17.2-5.** O Midtown Bank sempre tem dois caixas em serviço. Os clientes chegam para ser atendidos por um caixa em uma taxa média de 40 por hora. Um caixa precisa de uma média de dois minutos para atender um cliente. Quando os dois caixas estão ocupados, um cliente que chega junta-se a uma fila única para esperar por atendimento. A experiência demonstra que os clientes aguardam na fila em média um minuto antes de ser atendidos.

**(a)** Descreva por que esse é um sistema de filas.

**(b)** Determine as medidas básicas de desempenho — $W_q$, $W$, $L_q$ e $L$ — para esse sistema de filas. *Dica*: Não conhecemos as distribuições probabilísticas dos tempos entre atendimentos e tempos de atendimento para esse sistema de filas, de modo que precisaremos usar as relações entre essas medidas de desempenho para ajudá-lo a responder a estas perguntas.

**17.2-6.** Explique por que o fator de utilização $\rho$ para o atendente em um sistema de filas com um único atendente tem de ser igual a $1 - P_0$, em que $P_0$ é a probabilidade de ter 0 clientes no sistema.

**17.2-7.** São dados dois sistemas de filas, $Q_1$ e $Q_2$. A taxa média de chegada, a taxa média de atendimento por atendente ocupado e o número esperado de clientes em estado estável para $Q_2$ são o dobro dos valores correspondentes para $Q_1$. Façamos que $W_i$ = tempo de espera previsto em estado estável no sistema para $Q_i$, para $i = 1, 2$. Determine $W_2/W_1$.

**17.2-8.** Considere um sistema de filas com um único atendente com uma distribuição de tempo de atendimento *qualquer* e uma distribuição de tempos entre atendimentos *qualquer* (o modelo *GI*/*G*/1). Use somente definições básicas e as relações dadas na Seção 17.2 para verificar as seguintes relações gerais:

**(a)** $L = L_q + (1 - P_0)$.

**(b)** $L = L_q + \rho$.

**(c)** $P_0 = 1 - \rho$.

**17.2-9.** Demostre que

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right)$$

usando as definições estatísticas de $L$ e $L_q$ em termos de $P_n$.

**17.3-1.** Identifique os clientes e os atendentes no sistema de filas em cada uma das seguintes situações:

**(a)** O caixa em uma loja.

**(b)** Um posto de corpo de bombeiros.

**(c)** A cabine de pedágio em uma ponte.

**(d)** Uma loja de conserto de bicicletas.

**(e)** Um terminal marítimo.

**(f)** Um grupo de máquinas semi-automáticas designadas a um operador.

**(g)** O equipamento de manipulação de materiais em uma área de uma fábrica.

**(h)** Uma loja de tubos e conexões.

**(i)** Uma empreiteira atendendo pedidos sob encomenda.

**(j)** Um *pool* de datilógrafas.

**17.4-1.** Suponha que um sistema de filas tenha dois atendentes, uma distribuição de tempos entre atendimentos exponencial com uma média de duas horas e uma distribuição de tempos de atendimento exponencial com uma média de duas horas para cada um dos atendentes. Além disso, um cliente acaba de chegar ao meio-dia.

**(a)** Qual é a probabilidade de que a próxima chegada se dará (i) antes das 13 h, (ii) entre 13 e 14 h e (iii) após as 14 h?

**(b)** Suponha que não haja mais nenhuma chegada de cliente antes das 13 h. Qual é a probabilidade agora de que a próxima chegada venha a ocorrer entre 13 e 14 h?

**(c)** Qual é a probabilidade de que o número de chegadas entre 13 e 14 h será: (i) 0, (ii) 1 e (iii) 2 ou mais?

**(d)** Suponha que ambos os atendentes estejam com clientes às 13 h. Qual é a probabilidade de que *nenhum* dos clientes terá seu atendimento completado (i) antes das 14 h, (ii) antes das 13h10 e (iii) antes das 13h01 h?

**17.4-2.*** As tarefas a ser executadas em determinada máquina chegam de acordo com um *processo de entrada de Poisson* com uma taxa média de duas por hora. Suponha que a máquina quebre e exija uma hora para ser reparada. Qual é a probabilidade de que o número de tarefas novas que chegarão durante esse período seja de: (*a*) 0, (*b*) 2 e (*c*) 5 ou mais?

**17.4-3.** O tempo necessário para um mecânico consertar uma máquina tem uma distribuição exponencial com uma média de quatro horas. Entretanto, uma ferramenta especial reduziria essa média para duas horas. Se o mecânico consertar a máquina em menos de duas horas, ele receberá US$ 100; caso contrário, ele receberá US$ 80. Determine o aumento esperado no pagamento do mecânico por máquina consertada caso ele use a ferramenta especial.

**17.4-4.** Um sistema de filas com três atendentes possui um processo de chegada controlado que libera clientes a tempo de manter os atendentes sempre ocupados. Os tempos de atendimento têm uma distribuição exponencial com média 0,5.

Você observa o sistema de filas iniciando com todos os três atendentes começando a atender no instante $t = 0$. A seguir, você percebe que o primeiro término de atendimento ocorre no instante $t = 1$. Dadas essas informações, determine o tempo esperado após $t = 1$ até a ocorrência do próximo término de atendimento.

**17.4-5.** Um sistema de filas possui três atendentes com tempos de atendimento esperado de 20 minutos, 15 minutos e dez minutos. Os tempos de atendimento possuem uma distribuição exponencial. Cada um dos atendentes tem-se mantido ocupado com um cliente por cinco minutos. Determine o tempo restante esperado até que aconteça o próximo término de atendimento.

**17.4-6.** Considere um sistema de filas com dois tipos de clientes. Os clientes do tipo 1 chegam de acordo com um processo de Poisson com uma taxa média de 5 por hora. Os clientes do tipo 2 também chegam de acordo com um processo de Poisson com uma taxa média de 5 por hora. O sistema possui dois atendentes, ambos atendem os dois tipos de clientes. Para ambos os tipos, os tempos de atendimento possuem uma distribuição exponencial com uma média de dez minutos. O atendimento é feito segundo a regra na qual os primeiros que chegam serão os primeiros a ser atendidos.

**(a)** Qual é a distribuição probabilística (incluindo sua média) do tempo entre chegadas consecutivas de clientes de qualquer tipo?

**(b)** Quando determinado cliente do tipo 2 chega, ele encontra dois clientes do tipo 1 no processo de serem atendidos, porém nenhum outro cliente no sistema. Qual é a distribuição probabilística (incluindo sua média) do tempo de espera na fila desse cliente de tipo 2?

**17.4-7.** Considere um sistema de filas com dois atendentes em que todos os tempos de atendimento são independentes e identicamen-

te distribuídos de acordo com uma distribuição exponencial com uma média de dez minutos. O atendimento é fornecido segundo a regra na qual os primeiros que chegam serão os primeiros a ser atendidos. Quando determinado cliente chega, ele encontra os dois atendentes ocupados e ninguém esperando na fila.

**(a)** Qual é a distribuição probabilística (incluindo sua média e desvio-padrão) do tempo de espera na fila desse cliente?

**(b)** Determine o valor esperado e desvio-padrão do tempo de espera no sistema desse cliente.

**(c)** Suponha que esse cliente ainda esteja esperando na fila cinco minutos após sua chegada. Dada essa informação, como isso muda o valor esperado e o desvio-padrão do tempo de espera total no sistema desse cliente das respostas obtidas no item (*b*)?

**17.4-8.** Para cada uma das seguintes alternativas referentes a tempos de atendimento modelados pela distribuição exponencial, classifique a alternativa como verdadeira ou falsa e depois justifique sua resposta referindo-se a afirmações específicas (citando o número da página) no capítulo.

**(a)** O valor esperado e a variância dos tempos de atendimento são sempre iguais.

**(b)** A distribuição exponencial sempre fornece uma boa aproximação da distribuição de tempos de atendimento real quando cada cliente requer as mesmas operações de atendimento.

**(c)** Em uma instalação com $s$ atendentes, $s > 1$, com exatamente $s$ clientes já no sistema, uma nova chegada teria um tempo de espera previsto antes de ser atendida de $1/\mu$ unidades de tempo, em que $\mu$ é a taxa média de atendimento para cada atendente ocupado.

**17.4-9.** Assim como para a Propriedade 3 da distribuição exponencial, façamos que $T_1, T_2, \ldots, T_n$ sejam variáveis aleatórias exponenciais independentes com parâmetros $\alpha_1, \alpha_2, \ldots, \alpha_n$, respectivamente, e façamos que $U = \min\{T_1, T_2, \ldots, T_n\}$. Demonstre que a probabilidade de que determinada variável aleatória $T_j$ venha a ser a menor das $n$ variáveis aleatórias é

$$P\{T_j = U\} = \alpha_j \Big/ \sum_{i=1}^{n} \alpha_i, \qquad \text{para } j = 1, 2, \ldots, n.$$

*Dica*: $P\{T_j = U\} = \int_0^\infty P\{T_i > T_j \text{ para todo } i \neq j \,|\, T_j = t\}\alpha_j e^{-\alpha_j t} dt$.

**17.5-1.** Considere o processo de nascimento-e-morte com todos os $\mu_n = 2$ ($n = 1, 2, \ldots$), $\lambda_0 = 3$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_n = 0$ para $n = 3, 4, \ldots$.

**(a)** Mostre o diagrama de taxas.

**(b)** Calcule $P_0$, $P_1$, $P_2$, $P_3$ e $P_n$ para $n = 4, 5, \ldots$.

**(c)** Calcule $L$, $L_q$, $W$ e $W_q$.

**17.5-2.** Considere um processo de nascimento-e-morte com apenas três estados atingíveis (0, 1 e 2), para os quais as probabilidades de estado estável são, respectivamente, $P_0$, $P_1$ e $P_2$. As taxas de nascimento-e-morte são sintetizadas na seguinte tabela:

| Estado | Taxa de Nascimento | Taxa de Mortalidade |
|---|---|---|
| 0 | 1 | — |
| 1 | 1 | 2 |
| 2 | 0 | 2 |

**(a)** Construa o diagrama de taxas para esse processo de nascimento-e-morte.

**(b)** Desenvolva as equações de equilíbrio.

**(c)** Solucione essas equações para encontrar $P_0$, $P_1$ e $P_2$.

**(d)** Use as fórmulas gerais para o processo de nascimento-e-morte para calcular $P_0$, $P_1$ e $P_2$. Calcule também $L$, $L_q$, $W$ e $W_q$.

**17.5-3.** Considere o processo de nascimento-e-morte com as seguintes taxas médias. As taxas de nascimento são $\lambda_0 = 2$, $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1$ e $\lambda_n = 0$ para $n > 3$. As taxas de mortalidade são $\mu_1 = 3$, $\mu_2 = 4$, $\mu_3 = 1$ e $\mu_n = 2$ para $n > 4$.

**(a)** Construa o diagrama de taxas para esse processo de nascimento-e-morte.

**(b)** Desenvolva as equações de equilíbrio.

**(c)** Resolva essas equações para encontrar a distribuição probabilística de estado estável $P_0, P_1, \ldots$.

**(d)** Use as fórmulas genéricas para o processo de nascimento-e-morte para calcular $P_0, P_1, \ldots$. Calcule também $L$, $L_q$, $W$ e $W_q$.

**17.5-4.** Considere o processo de nascimento-e-morte com todos $\lambda_n = 2$ ($n = 0, 1, \ldots$), $\mu_1 = 2$ e $\mu_n = 4$ para $n = 2, 3, \ldots$.

**(a)** Exiba o diagrama de taxas.

**(b)** Calcule $P_0$ e $P_1$. A seguir, forneça uma expressão genérica para $P_n$ em termos de $P_0$ para $n = 2, 3, \ldots$.

**(c)** Considere um sistema de filas com dois atendentes que se encaixe nesse processo. Qual é a taxa média de chegada para esse sistema de filas? Qual é a taxa média de atendimento para cada um dos atendentes quando ele se encontra ocupado atendendo clientes?

**17.5-5.*** Um posto possui uma bomba de gasolina. Carros querendo gasolina chegam de acordo com um processo de Poisson a uma taxa média de 15 por hora. Entretanto, se a bomba estiver sendo usada, esses possíveis clientes poderão *se recusar* (ir para outro posto de gasolina). Particularmente, se tivermos $n$ carros já no posto, a probabilidade de chegar um possível cliente que vai se recusar é de $n/3$ para $n = 1, 2, 3$. O tempo necessário para abastecer um carro tem uma distribuição exponencial com uma média de quatro minutos.

**(a)** Construa o diagrama de taxas para esse sistema de filas.

**(b)** Desenvolva as equações de equilíbrio.

**(c)** Resolva essas equações para encontrar uma distribuição probabilística de estado estável do número de carros que se encontram no posto. Verifique que essa solução é a mesma daquela dada pela solução geral para o processo de nascimento-e-morte.

**(d)** Encontre o tempo de espera previsto (incluindo atendimento) para aqueles carros que permanecem.

**17.5-6.** Um técnico de manutenção tem a tarefa de manter duas máquinas em funcionamento. O tempo que uma máquina trabalha antes de quebrar possui uma distribuição exponencial com uma média de dez horas. O tempo gasto então pelo técnico de manutenção para reparar a máquina possui uma distribuição exponencial com uma média de oito horas.

**(a)** Demonstre que esse processo se ajusta ao processo de nascimento-e-morte definindo os estados, especificando os valores de $\lambda_n$ e $\mu_n$ e então construindo o diagrama de taxas.

**(b)** Calcule o $P_n$.

**(c)** Calcule $L$, $L_q$, $W$ e $W_q$.

**(d)** Determine a proporção de tempo que o técnico de manutenção está ocupado.

**(e)** Determine a proporção de tempo que dada máquina está operando.

**(f)** Volte ao exemplo quase idêntico da *cadeia de Markov de tempo contínuo* dado no final da Seção 16.8. Descreva a relação entre cadeias de Markov de tempo contínuo e o processo de nascimento-e-morte que permite que ambos sejam aplicados a esse mesmo problema.

**17.5-7.** Considere um sistema de filas com um único atendente em que tempos entre atendimentos possuem uma distribuição exponencial com parâmetro $\lambda$ e tempos de atendimento têm uma distribuição exponencial com parâmetro $\mu$. Além disso, clientes *desistem* (deixam o sistema de filas sem ser atendidos) caso seus tempos de espera na fila acabem ficando muito grandes. Em particular, suponha que o tempo que cada cliente está disposto a esperar na fila antes de desistir tenha uma distribuição exponencial com média $1/\theta$.

**(a)** Construa o diagrama de taxas para esse sistema de filas.

**(b)** Desenvolva as equações de equilíbrio.

**17.5-8.*** Determinada mercearia pequena possui um único caixa em tempo integral. Os clientes chegam "aleatoriamente" ao caixa (isto é, um processo de entrada de Poisson) a uma taxa média de 30 por hora. Quando há apenas um cliente na fila, ele é processado somente pelo caixa, com um tempo de atendimento esperado de 1,5 minuto. Entretanto, o ajudante de estoque recebeu ordens de sempre que houvesse mais de um cliente na fila, ele deve ajudar o caixa, empacotando as mercadorias. Essa ajuda reduz o tempo esperado para processar um cliente a um minuto. Em ambos os casos, a distribuição de tempo de atendimento é exponencial.

**(a)** Construa o diagrama de taxas para esse sistema de filas.

**(b)** Qual é a distribuição probabilística em estado estável do número de clientes no caixa?

**(c)** Derive $L$ para esse sistema. *Dica*: Volte à derivação de $L$ para o modelo *M/M/*1 no início da Seção 17.6. Use essas informações para determinar $L_q$, $W$ e $W_q$.

**17.5-9.** Um departamento possui um operador de processador de texto. Documentos produzidos pelo departamento são entregues para processamento de acordo com um processo de Poisson com tempos entre atendimento esperado de 20 minutos. Quando o operador tem apenas um documento para processar, o tempo de processamento esperado é de 15 minutos. Quando ele tem mais de um documento, então ajuda para edição que se encontra disponível reduz o tempo de processamento esperado para cada documento a dez minutos. Em ambos os casos, os tempos de processamento possuem uma distribuição exponencial.

**(a)** Construa o diagrama de taxas para esse sistema de filas.

**(b)** Encontre a distribuição de estado estável do número de documentos que o operador recebeu, mas não completou ainda.

**(c)** Derive $L$ para esse sistema. *Dica:* Volte à derivação de $L$ para o modelo *M/M/*1 no início da Seção 17.6. Use essa informação para determinar $L_q$, $W$ e $W_q$.

**17.5-10.** Clientes chegam em um sistema de filas de acordo com um processo de Poisson a uma taxa média de chegada de 2 clientes por minuto. O tempo de atendimento possui uma distribuição exponencial com uma média de um minuto. Um número ilimitado de atendentes está disponível conforme a necessidade de modo que os clientes jamais têm de esperar para ser atendidos. Calcule a probabilidade de estado estável de que exatamente um cliente se encontre no sistema.

**17.5-11.** Suponha que um sistema de filas com um único atendente se ajuste a todas as hipóteses do processo de nascimento-e-morte, *exceto* que clientes sempre chegam em *pares*. A taxa média de chegada é de dois pares por hora (quatro clientes por hora) e a taxa média de atendimento (quando o atendente estiver ocupado) é de cinco clientes por hora.

**(a)** Construa o diagrama de taxas para esse sistema de filas.

**(b)** Desenvolva as equações de equilíbrio.

**(c)** Para fins de comparação, mostre o diagrama de taxas para o sistema de filas correspondente que se ajusta completamente ao processo de nascimento-e-morte, isto é, onde clientes chegam *individualmente* em uma taxa média de quatro por hora.

**17.5-12.** Considere um sistema de filas com um único atendente com uma fila finita capaz de reter no máximo dois clientes, *excluindo* qualquer um que esteja sendo atendido. O atendente é capaz de fornecer *atendimento em lote* a dois clientes simultaneamente, onde o tempo de atendimento tem uma distribuição exponencial com uma média de uma unidade de tempo independentemente do número que estiver sendo atendido. Sempre que a fila não estiver cheia, os clientes chegam individualmente de acordo com um processo de Poisson a uma taxa média de 1 por unidade de tempo.

**(a)** Suponha que o atendente *deva* atender dois clientes simultaneamente. Portanto, se o atendente estiver ocioso quando somente um cliente estiver no sistema, o atendente tem de esperar por outra chegada antes de iniciar o atendimento. Formule o modelo de filas como uma cadeia de Markov de tempo contínuo definindo os estados e depois construindo o diagrama de taxas. Forneça as equações de equilíbrio, mas não as resolva.

**(b)** Suponha agora que o tamanho do lote para um atendimento seja 2 somente se dois clientes estiverem na fila quando o atendente terminar o atendimento anterior. Portanto, se o atendente estiver ocioso quando apenas um cliente estiver no sistema, o atendente deve atender esse único cliente e quaisquer chegadas subseqüentes deverão aguardar na fila até que o atendimento seja completado para esse cliente. Formule o modelo de filas resultante como uma cadeia de Markov de tempo contínuo definindo os estados e depois construindo o diagrama de taxas. Forneça as equações de equilíbrio, mas não as resolva.

**17.5-13.** Considere um sistema de filas com duas classes de clientes, dois escriturários atendendo e *nenhuma fila.* Possíveis clientes de cada classe chegam de acordo com um processo de Poisson, com uma taxa média de chegada de dez clientes por hora para a classe 1 e 5 clientes por hora para a classe 2, porém essas chegadas são perdidas para o sistema, caso elas não possam ser atendidas imediatamente.

Cada cliente da classe 1 que entra no sistema será atendido por um dos escriturários que estiver livre, onde os tempos de atendimento possuem uma distribuição exponencial com uma média de cinco minutos.

Cada cliente de classe 2 que entrar no sistema requer o *uso simultâneo de ambos os escriturários* (os dois escriturários trabalham juntos com um único atendente), onde os tempos de atendimento possuem uma distribuição exponencial com uma média de cinco minutos. Portanto, um cliente que chega desse tipo seria perdido para o sistema a menos que ambos os escriturários estejam livres para começar a atender imediatamente.

**(a)** Formule o modelo de filas como uma cadeia de Markov de tempo contínuo definindo os estados e construindo o diagrama de taxas.

**(b)** Agora, descreva como a formulação no item (*a*) pode ser adequada no formato do processo de nascimento-e-morte.
**(c)** Use os resultados para o processo de nascimento-e-morte para calcular a distribuição conjunta de estado estável do número de clientes de cada classe no sistema.
**(d)** Para cada uma das duas classes de clientes, qual é a parcela de chegadas esperada que se encontram incapazes de entrar no sistema?

**17.6-1.*** A 4M Company possui um único torno-revólver como principal máquina de usinagem em seu chão de fábrica. As tarefas chegam nessa máquina de acordo com um processo de Poisson em uma taxa média de 2 por dia. O tempo de processamento para realizar cada tarefa tem uma distribuição exponencial com uma média de $\frac{1}{4}$ dia. Como as tarefas são volumosas, aquelas que não estão sendo trabalhadas no momento estão sendo armazenadas em uma sala a certa distância da máquina. Entretanto, para poupar tempo na produção das tarefas, o gerente de produção está propondo adicionar espaço de armazenagem para produtos em fabricação suficiente próximo ao torno-revólver para acomodar três tarefas além daquela que está sendo processada no momento. Tarefas em excesso continuarão a ser armazenadas temporariamente na sala distante. Segundo essa proposta, que proporção de tempo esse espaço de armazenagem próximo ao torno-revólver é adequado para acomodar todas as tarefas em espera?
**(a)** Use fórmulas disponíveis para calcular sua resposta.
T **(b)** Use o gabarito de Excel correspondente para obter as probabilidades necessárias para responder a pergunta.

**17.6-2.** Clientes chegam em um sistema de filas com um único atendente de acordo com um processo de Poisson em uma taxa média de 10 por hora. Se o atendente trabalhar continuamente, o número de clientes que podem ser atendidos em uma hora tem uma distribuição de Poisson com média 15. Determine a proporção de tempo durante o qual ninguém está esperando para ser atendido.

**17.6-3.** Considere o modelo *M*/*M*/1, com $\lambda < \mu$.
**(a)** Determine a probabilidade de estado estável de que o tempo de espera real no sistema de um cliente não é maior que o tempo de espera previsto no sistema, isto é, $P\{\mathcal{W} > W\}$.
**(b)** Determine a probabilidade de estado estável de que o tempo de espera real na fila de um cliente não seja maior que o tempo de espera previsto na fila, isto é, $P\{\mathcal{W}_q > W_q\}$.

**17.6-4.** Verifique as seguintes relações para um sistema de filas *M*/*M*/1:

$$\lambda = \frac{(1 - P_0)^2}{W_q P_0}, \qquad \mu = \frac{1 - P_0}{W_q P_0}.$$

**17.6-5.** É necessário determinar quanto espaço de armazenagem para itens em fabricação alocar a uma determinada máquina em uma nova fábrica. As tarefas chegam nessa máquina de acordo com um processo de Poisson com uma taxa média de 3 por hora e o tempo necessário para realizar o trabalho necessário tem uma distribuição exponencial com uma média de 0,25 hora. Toda vez que os tempos de espera exigirem mais espaço de armazenagem para itens em fabricação do que foi alocado, as tarefas em excesso são armazenadas temporariamente em um local menos conveniente. Se cada tarefa exigir 1 $m^2$ de espaço no chão de fábrica, enquanto estiver sendo armazenada temporariamente junto à máquina durante a fabricação, quanto espaço deve ser fornecido para acomodar todas as tarefas em espera: (*a*) 50% do tempo, (*b*) 90% do tempo e (*c*) 99% do tempo? Obtenha uma expressão analítica para responder essas três perguntas. *Dica:* A soma de uma série geométrica é

$$\sum_{n=0}^{N} x^n = \frac{1 - x^{N+1}}{1 - x}.$$

**17.6-6.** Considere as seguintes alternativas sobre um sistema de filas *M*/*M*/1 e seu fator de utilização $\rho$. Classifique cada uma das alternativas como verdadeira ou falsa e então justifique sua resposta.
**(a)** A probabilidade de que um cliente tenha de esperar antes de ser atendido é proporcional a $\rho$.
**(b)** O número de clientes esperado no sistema é proporcional a $\rho$.
**(c)** Se $\rho$ tiver sido aumentado de $\rho = 0{,}9$ a $\rho = 0{,}99$, o efeito de qualquer outro aumento em $\rho$ sobre $L$, $L_q$, $W$ e $W_q$ será relativamente pequeno desde que $\rho < 1$.

**17.6-7.** Clientes chegam em um sistema de filas com um único atendente de acordo com um processo de Poisson com um tempo esperado entre atendimentos de 25 minutos. Tempos de atendimento possuem uma distribuição exponencial com média de 30 minutos.

Classifique cada uma das seguintes alternativas sobre esse sistema como verdadeira ou falsa e, a seguir, justifique sua resposta.
**(a)** O atendente certamente estará ocupado para sempre após o primeiro cliente chegar.
**(b)** A fila crescerá sem limites.
**(c)** Se for acrescentado um segundo atendente com a mesma distribuição de tempos de atendimento, o sistema pode alcançar uma condição de estado estável.

**17.6-8.** Para cada uma das seguintes alternativas sobre um sistema de filas *M*/*M*/1, classifique cada uma das seguintes alternativas sobre esse sistema como verdadeira ou falsa e, a seguir, justifique sua resposta referindo-se a afirmações específicas (citando o número da página) no capítulo.
**(a)** O tempo de espera no sistema possui uma distribuição exponencial.
**(b)** O tempo de espera na fila possui uma distribuição exponencial.
**(c)** O tempo condicional de espera no sistema, dado o número de clientes já no sistema, possui uma distribuição de Erlang (distribuição gama).

**17.6-9.** A Friendly Neighbor Grocery Store possui um terminal de caixa com um caixa em tempo integral. Clientes chegam aleatoriamente no caixa a uma taxa média de 30 por hora. A distribuição de tempo de atendimento é exponencial, com uma média de 1,5 minuto. Essa situação resultou, ocasionalmente, em longas filas e reclamações por parte dos clientes. Portanto, como não há espaço para um segundo terminal de caixa, o gerente está considerando a alternativa de contratar outra pessoa para ajudar o caixa empacotando as mercadorias. Essa ajuda reduziria o tempo esperado para processar um cliente para um minuto, porém a distribuição ainda seria exponencial.

O gerente gostaria de ter a porcentagem de tempo onde há mais de dois clientes no caixa abaixo de 25%. Ele também gostaria de que não mais de 5% dos clientes tivessem de esperar na fila pelo menos cinco minutos antes de ser atendido ou pelo menos sete minutos antes de terminar o atendimento.
**(a)** Use as fórmulas para o modelo *M*/*M*/1 para calcular $L$, $W$, $W_q$, $L_q$, $P_0$, $P_1$ e $P_2$ para o modo de operação atual. Qual é a probabilidade de ter mais de dois clientes no caixa?

T **(b)** Use o gabarito em Excel para esse modelo para verificar suas respostas no item (*a*). Encontre também a probabilidade de que o tempo de espera antes do atendimento exceda cinco minutos e a probabilidade de que o tempo de espera antes de terminar o atendimento exceda sete minutos.
**(c)** Repita o item (*a*) para a alternativa considerada pelo gerente.
**(d)** Repita o item (*b*) para essa alternativa.
**(e)** Que abordagem o gerente deveria usar para satisfazer seus critérios o mais próximo possível?

T **17.6-10.** O Centerville International Airport possui duas pistas, uma usada exclusivamente para levantar vôo e a outra exclusiva para aterrissagens. Os aviões chegam no espaço aéreo de Centerville para solicitar instruções de pouso de acordo com um processo de Poisson em uma taxa média de 10 por hora. O tempo necessário para um avião pousar após receber autorização para fazê-lo tem uma distribuição exponencial com uma média de três minutos e esse processo tem de ser completado antes de dar autorização para pouso para outro avião. Aviões aguardando autorização devem circular pelo aeroporto.

A Administração Federal da Aviação tem uma série de critérios referentes ao nível de segurança de congestionamento de aviões aguardando para pousar. Esses critérios dependem de uma série de fatores referentes ao aeroporto envolvido, como o número de pistas disponíveis para aterrissagem. Para o Centerville, os critérios são: (1) o número médio de aviões aguardando para receber autorização para pouso não deve exceder 1, (2) 95% do tempo, o número real de aviões aguardando para receber autorização para pouso não deve exceder 4, (3) para 99% dos aviões, o tempo gasto circulando o aeroporto antes de receber autorização para pouso não deve exceder 30 minutos (já que exceder esse período normalmente exigiria o redirecionamento do avião para outro aeroporto para um pouso de emergência antes que seu combustível acabe).
**(a)** Avalie em que nível esses critérios estão sendo satisfeitos no momento.
**(b)** Uma importante companhia aérea considera a possibilidade de adicionar esse aeroporto como um de seus principais terminais. Isso aumentaria a taxa média de chegada a 15 aviões por hora. Avalie em que nível os critérios anteriores seriam satisfeitos se isso acontecesse.
**(c)** Para atrair mais negócios [inclusive a importante companhia aérea mencionada no item (*b*)], a gerência do aeroporto considera uma segunda pista para pouso. Estima-se que esta aumentaria finalmente a taxa média de chegada para 25 aviões por hora. Avalie em que nível os critérios anteriores seriam satisfeitos caso isso acontecesse.

T **17.6-11.** O Security & Trust Bank emprega quatro caixas para atender a seus clientes. Os clientes chegam de acordo com um processo de Poisson a uma taxa média de 2 por minuto. Entretanto, o negócio está crescendo e a gerência projeta que a taxa média de chegada será 3 por minuto daqui a um ano. O tempo de transação entre o caixa e o cliente tem uma distribuição exponencial com média de um minuto.

A gerência estabeleceu as seguintes diretrizes para um nível satisfatório de atendimento aos clientes. O número médio de clientes esperando na fila para ser atendidos não deve exceder 1. Pelo menos 95% do tempo, o número de clientes esperando na fila não deve exceder 5. Para pelo menos 95% dos clientes, o tempo gasto na fila esperando para ser atendido não deve ultrapassar cinco minutos.
**(a)** Use o modelo *M/M/s* para determinar o nível em que essas diretrizes estão sendo satisfeitas.
**(b)** Avalie em que nível as diretrizes serão satisfeitas daqui a um ano caso não seja feita nenhuma alteração no número de caixas.
**(c)** Determine quantos caixas serão necessários daqui a um ano para atender completamente a essas diretrizes.

**17.6-12.** Considere o modelo *M/M/s*.
T **(a)** Suponha que haja um atendente e o tempo de atendimento esperado seja exatamente igual a um minuto. Compare $L$ para os casos nos quais a taxa média de chegada é de 0,5, 0,9 e 0,99 cliente por minuto, respectivamente. Faça o mesmo para $L_q$, $W$, $W_q$ e $P\{\mathcal{W} > 5\}$. Que conclusões você tira sobre o impacto de aumentar o fator de utilização $\rho$ de valores pequenos (por exemplo, $\rho = 0{,}5$) para valores bem altos (por exemplo, $\rho = 0{,}9$) e depois para valores maiores ainda, próximos a 1 (por exemplo, $\rho = 0{,}99$)?
**(b)** Suponha agora que haja dois atendentes e o tempo de atendimento esperado seja exatamente de dois minutos. Siga as instruções para o item (*a*).

T **17.6-13.** Considere o modelo *M/M/s* com taxa média de chegada de dez clientes por hora e um tempo de atendimento esperado de cinco minutos. Use o gabarito em Excel para esse modelo para obter e imprimir as diversas medidas de desempenho (com $t = 10$ e $t = 0$, respectivamente, para as duas probabilidades de tempo de espera) quando o número de atendentes for 1, 2, 3, 4 e 5. Depois, para cada um dos seguintes critérios possíveis para um nível satisfatório de atendimento (em que a unidade de tempo é de 1 minuto), use os resultados impressos para determinar quantos atendentes são necessários para satisfazer esse critério.
**(a)** $L_q \leq 0{,}25$
**(b)** $L \leq 0{,}9$
**(c)** $W_q \leq 0{,}1$
**(d)** $W \leq 6$
**(e)** $P\{\mathcal{W}_q > 0\} \leq 0{,}01$
**(f)** $P\{\mathcal{W} > 10\} \leq 0{,}2$
**(g)** $\sum_{n=0}^{s} P_n \geq 0.95$

**17.6-14.** Um posto de gasolina com apenas uma bomba adota a seguinte política: se um cliente tiver de esperar, o preço é de US$ 1 por litro; se não tiver de esperar, o preço será de US$ 1,20 por litro. Os clientes chegam de acordo com um processo de Poisson com taxa média de 15 por hora. Os tempos de atendimento na bomba têm uma distribuição exponencial com média de três minutos. Os clientes que chegam sempre aguardam até eles finalmente poderem adquirir o combustível. Determine o preço esperado da gasolina por litro.

**17.6-15.** É fornecido um sistema de fila *M*/*M*/1 com taxa média de chegada $\lambda$ e taxa média de atendimento $\mu$. Um cliente que chega recebe $n$ dólares, caso $n$ clientes já se encontrarem no sistema. Determine o custo esperado em dólares por cliente.

**17.6-16.** A Seção 17.6 fornece as seguintes equações para o modelo *M*/*M*/1:

(1) $$P\{\mathcal{W} > t\} = \sum_{n=0}^{\infty} P_n P\{S_{n+1} > t\}.$$

(2) $$\mathrm{P}\{\mathcal{W} > \mathrm{t}\} = \mathrm{e}^{-\mu(1-\rho)t}.$$

Demonstre que a Eq. (1) se reduz algebricamente à Eq. (2). *Dica:* Use diferenciação, álgebra e integração.

**17.6-17.** Obtenha $W_q$ diretamente para os seguintes casos desenvolvendo e reduzindo uma expressão análoga à Eq. (1) no Problema 17.6-16. *Dica:* Use o tempo de espera *condicional* previsto na fila dado que uma chegada aleatória encontra *n* clientes já no sistema.

**(a)** O modelo *M/M/*1

**(b)** O modelo *M/M/s*

T **17.6-18.** Considere um sistema de filas *M*/*M*/2 com $\lambda = 4$ e $\mu = 3$. Determine a taxa média na qual os términos de atendimento ocorrem durante os períodos nos quais os clientes se encontram aguardando na fila.

T **17.6-19.** Dado um sistema de filas *M*/*M*/2 com $\lambda = 4$ por hora e $\mu = 6$ por hora. Determine a probabilidade de que um cliente que chega vá aguardar mais de 30 minutos na fila, dado que pelo menos dois clientes já se encontrem no sistema.

**17.6-20.*** Na Cia. de Seguros Blue Chip Life, as funções de depósito e de retirada associadas a certo produto de investimento são destinadas separadamente a dois escriturários, Clara e Clarence. Os comprovantes de depósito chegam aleatoriamente (um processo de Poisson) na mesa de Clara a uma taxa média de 16 por hora. Comprovantes de retirada chegam aleatoriamente (um processo de Poisson) na mesa de Clarence a uma taxa média de 14 por hora. O tempo necessário para processar qualquer uma das transações apresenta uma distribuição exponencial com média de três minutos. Para reduzir o tempo de espera previsto no sistema tanto para os comprovantes de depósito quanto de retirada, o departamento atuarial fez as seguintes recomendações: (1) Treinar cada escriturário para lidar com depósitos e retiradas e (2) colocar os recibos de depósito bem como os de retirada em uma única fila que fosse acessada por ambos os escriturários.

**(a)** Determine o tempo de espera previsto no sistema segundo os procedimentos atuais para cada tipo de comprovante. A seguir, combine esses resultados para calcular o tempo de espera previsto no sistema para uma chegada aleatória de cada tipo de envelope.

T **(b)** Se as recomendações forem adotadas, determine o tempo de espera previsto no sistema para chegada dos envelopes.

T **(c)** Suponha agora que a adoção das recomendações resultasse em um ligeiro aumento no tempo de processamento esperado. Use o gabarito em Excel para o modelo *M/M/s* para determinar por tentativa e erro o tempo de processamento esperado (dentro de 0,001 hora) que faria o tempo de espera previsto no sistema para uma chegada aleatória ser o mesmo segundo os procedimentos atuais e as novas recomendações.

**17.6-21.** A People's Software Company acaba de instalar um *call center* para fornecer assistência técnica para o seu novo pacote de software. Dois técnicos estão atendendo as ligações, nas quais o tempo necessário para cada um dos técnicos responder às perguntas de um cliente apresenta uma distribuição exponencial com média de oito minutos. Os telefonemas estão chegando de acordo com um processo de Poisson a uma taxa média de 10 por hora.

Espera-se que no ano que vem a taxa média de chegada de ligações caia para 5 por hora, de modo que o plano seja reduzir o número de técnicos para somente um.

T **(a)** Supondo que $\mu$ continuará a ser 7,5 ligações por hora para o sistema de filas do próximo ano, determine $L$, $L_q$, $W$ e $W_q$ tanto para o sistema atual como para o sistema do próximo ano. Para cada uma dessas quatro medidas de desempenho, que sistema leva ao menor valor?

**(b)** Suponha agora que $\mu$ será ajustável quando o número de técnicos for reduzido para um. Encontre algebricamente o valor de $\mu$ que resultaria no mesmo valor de $W$ daquele do atual sistema.

**(c)** Repita o item (*b*) com $W_q$ em vez de $W$.

**17.6-22.** Considere uma generalização do modelo *M/M/*1 em que o atendente precisa "se aquecer" no começo de um período movimentado e, portanto, atende o primeiro cliente de um período movimentado em uma velocidade menor que para os demais clientes. Particularmente, se um cliente que chega encontrar o atendente ocioso, o cliente passa por um tempo de atendimento com uma distribuição exponencial com parâmetro $\mu_1$. Entretanto, se um cliente que chega encontrar o atendente ocupado, esse cliente junta-se à fila e, posteriormente, passa por um tempo de atendimento com distribuição exponencial com parâmetro $\mu_2$, em que $\mu_1 < \mu_2$. Os clientes chegam de acordo com um processo de Poisson com taxa média $\lambda$.

**(a)** Formule esse modelo como uma cadeia de Markov de tempo contínuo definindo os estados e construindo o diagrama de taxas de acordo.

**(b)** Desenvolva as equações de equilíbrio.

**(c)** Suponha que sejam especificados valores numéricos para $\mu_1$, $\mu_2$ e $\lambda$ e que $\lambda < \mu_2$ (portanto, existe uma distribuição de estado estável). Já que esse modelo possui um número de estados infinito, a distribuição de estado estável é a solução simultânea de um número infinito de equações de equilíbrio (além da equação que a soma de probabilidades é igual a 1). Suponha que não seja possível obter essa solução analiticamente e, portanto, você deseja usar um computador para resolver o modelo numericamente. Considerando que seja impossível resolver um número infinito de equações numericamente, descreva brevemente o que ainda poderia ser feito com essas equações para obter uma aproximação de uma distribuição de estado estável. Sob que circunstâncias essa aproximação será basicamente exata?

**(d)** Dado que a distribuição de estado estável tenha sido obtida, forneça expressões explícitas para calcular $L$, $Lq$, $W$ e $Wq$.

**(e)** Dado essa distribuição de estado estável, desenvolva uma expressão para $P\{\mathcal{W} > t\}$ que seja análoga à Eq. (1) do Problema 17.6-16.

**17.6-23.** Para cada um dos modelos a seguir, escreva as equações de equilíbrio e demonstre que elas são satisfeitas pela solução dada na Seção 17.6 para a distribuição de estado estável do número de clientes no sistema.

**(a)** O modelo *M/M/*1.

**(b)** A variante de fila finita do modelo *M/M/*1, com $K = 2$.

**(c)** A variante finita da população solicitante do modelo *M/M/*1, com $N = 2$.

T **17.6-24.** Considere um sistema telefônico com três linhas. As ligações chegam de acordo com um processo de Poisson a uma taxa média de 6 por hora. A duração de cada ligação apresenta uma distribuição exponencial com média de 15 minutos. Se todas as

linhas estiverem ocupadas, as ligações serão colocadas em estado de espera até que alguma linha seja liberada.

**(a)** Imprima as medidas de desempenho fornecidas pelo gabarito em Excel para esse sistema de filas (com $t = 1$ hora e $t = 0$, respectivamente, para as duas probabilidades de tempo de espera).

**(b)** Use o resultado impresso fornecendo $P\{\mathcal{W}_q > 0\}$ para identificar a probabilidade de estado estável de que uma ligação será atendida imediatamente (não será colocada em espera). A seguir, verifique essa probabilidade usando os resultados impressos para $P_n$.

**(c)** Use os resultados impressos para identificar a distribuição probabilística de estado estável do número de ligações em espera.

**(d)** Imprima as novas medidas de desempenho caso as ligações que chegam sejam perdidas sempre que todas as linhas estiverem ocupadas. Use esses resultados para identificar a probabilidade de estado estável de que uma ligação que chega seja perdida.

**17.6-25.*** Janet planeja abrir um pequeno lava-rápido e ela tem de decidir sobre quanto espaço deverá ser disponibilizado para os carros que estão esperando sua vez. Ela estima que os clientes cheguem aleatoriamente (isto é, um processo de entrada de Poisson) com taxa média de 1 a cada quatro minutos, a menos que a área de espera esteja cheia, em cujo caso os clientes que chegam levariam seus carros em outro lugar. O tempo que pode ser atribuído para lavar um carro apresenta uma distribuição exponencial com média de três minutos. Compare a fração esperada de possíveis clientes que serão *perdidos* em razão do espaço de espera inadequado caso: (*a*) não exista nenhuma vaga (não inclui o carro que está sendo lavado), (*b*) duas vagas e (*c*) quatro vagas disponíveis.

**17.6-26.** Considere a variante de fila finita do modelo *M/M/s*. Derive a expressão para $L_q$ dada na Seção 17.6 para esse modelo.

**17.6-27.** Para a variante da fila finita do modelo *M/M/*1, desenvolva uma expressão análoga à Eq. (1) do Problema 17.6-16 para as seguintes probabilidades:

**(a)** $P\{\mathcal{W} > t\}$.

**(b)** $P\{\mathcal{W}_q > t\}$.

*Dica:* Chegadas podem ocorrer somente quando o sistema não estiver cheio e, portanto, a probabilidade de que uma chegada aleatória já encontre $n$ clientes lá é $P_n/(1 - P_K)$.

**17.6-28.** George planeja abrir uma cabina para revelação de fotos (*drive-through*) com um único posto de atendimento que ficará aberto aproximadamente 200 horas por mês em uma movimentada área comercial. Existe espaço disponível para aluguel de vagas para *drive-through* por US$ 200 mensais por vaga (espaço temporário ocupado por carro, disposto em fila). George precisa decidir quantas vagas devem ser fornecidas para seus clientes.

Excluindo esse custo de aluguel, George acredita que terá um lucro médio de US$ 4 por cliente atendido (nada para quando o filme for deixado e US$ 8 quando forem retiradas as fotografias reveladas). Ele também estima que os clientes chegarão de forma aleatória (um processo de Poisson) a uma taxa média de 20 por hora, embora aqueles que encontrem a fila do *drive-through* cheia serão obrigados a desistir. Metade dos clientes que encontram a fila cheia queriam deixar o filme e a outra metade queria pegar suas fotos reveladas. A metade que queria deixar filme acabará indo fazer isso em outra loja. A outra metade dos clientes que encontram a fila cheia não será perdida, pois tentará pegar suas fotos em outra oportunidade. George supõe que o tempo necessário para atender a um cliente terá uma distribuição exponencial com média de dois minutos.

T **(a)** Encontre $L$ e a taxa média nas quais os clientes são perdidos quando o número de vagas fornecidas for 2, 3, 4 e 5.

**(b)** Calcule $W$ a partir de $L$ para os casos considerados no item (*a*).

**(c)** Use os resultados do item (*a*) para calcular o decréscimo na taxa média na qual os clientes são perdidos quando o número de vagas fornecidas é aumentado de 2 para 3, de 3 para 4 e de 4 para 5. A seguir, calcule o aumento no lucro esperado por hora (excluindo os custos de aluguel) para cada um dos três casos.

**(d)** Compare os aumentos no lucro esperado encontrado no item (*c*) com o custo por hora de aluguel para cada vaga de carro. Que conclusão pode ser tirada sobre o número de vagas de carro que George deveria disponibilizar?

**17.6-29.** Na Forrester Manufacturing Company, foi atribuída a um técnico de manutenção a responsabilidade de fazer a manutenção de três máquinas. Para cada máquina, a distribuição probabilística do tempo em funcionamento antes de ocorrer uma quebra é exponencial, com média de nove horas. O tempo de reparo também apresenta uma distribuição exponencial, com média de duas horas.

**(a)** Qual modelo de filas se encaixa nesse sistema de filas?

T **(b)** Use esse modelo de filas para encontrar a distribuição probabilística do número de máquinas que não estão operando e a média dessa distribuição.

**(c)** Use essa média para calcular o tempo esperado entre a quebra de uma máquina e o término do reparo dessa máquina.

**(d)** Qual é a fração esperada de tempo que o técnico de manutenção ficará ocupado?

T **(e)** Como aproximação grosseira, suponha que a população solicitante seja infinita e que quebras de máquina ocorram aleatoriamente a uma taxa média de 3 a cada nove horas. Compare o resultado do item (*b*) com aquele obtido por fazer essa aproximação usando: (i) o modelo *M/M/s* e (ii) a variante de fila finita do modelo *M/M/s* com $K = 3$.

T **(f)** Repita o item (*b*) quando o segundo técnico é disponibilizado para reparar uma segunda máquina sempre que mais de uma das três máquinas precisar de reparo.

**17.6-30.** Reconsidere o processo de nascimento-e-morte específico descrito no Problema 17.5-1.

**(a)** Identifique um modelo de filas (e os valores de seus parâmetros) na Seção 17.6 que se encaixa nesse processo.

T **(b)** Use o gabarito em Excel correspondente para obter respostas para os itens (*b*) e (*c*) do Problema 17.5-1.

T **17.6-31.*** A Dolomite Corporation planeja construir uma nova fábrica. Foram alocadas 12 máquinas semi-automáticas a um departamento. Um pequeno número (ainda a ser determinado) de operadores será contratado para fornecer às máquinas o atendimento ocasional necessário (carregamento, descarregamento, ajuste, preparação e assim por diante). É preciso decidir agora como organizar os operadores para fazer isso. A alternativa 1 é alocar cada operador às suas próprias máquinas. A alternativa 2 é fazer um *pool* de operadores de modo que qualquer operador ocioso possa pegar a próxima máquina precisando de atendimento. A alternativa 3 é combinar os operadores em uma única equipe que trabalhará junta em qualquer máquina precisando de atendimento.

Supõe-se que o tempo em operação (tempo entre completar um atendimento e aquele de a máquina precisar de atendimento novamente) de cada máquina tenha uma distribuição exponencial, com média de 150 minutos. Supõe-se que o tempo de atendimento tenha uma distribuição exponencial, com média de 15 minutos (para as alternativas 1 e 2) ou 15 minutos divididos pelo número de operadores na equipe (para a alternativa 3). Para o departamento atingir a taxa de produção exigida, as máquinas têm de estar operando, em média, pelo menos 89% do tempo.

**(a)** Para a alternativa 1, qual é o número máximo de máquinas que pode ser alocado a um operador ainda mantendo a taxa de produção necessária? Qual é a utilização resultante de cada operador?

**(b)** Para a alternativa 2, qual é o número mínimo de operadores necessário para alcançar a taxa de produção exigida? Qual é a utilização resultante dos operadores?

**(c)** Para a alternativa 3, qual é o tamanho mínimo da equipe necessário para alcançar a taxa de produção necessária? Qual é a utilização resultante da equipe?

**17.6-32.** Uma ferramentaria possui três máquinas idênticas que estão sujeitas a falha de certo tipo. Portanto, é disponibilizado um sistema de manutenção para executar a operação de manutenção (recarga) necessária para uma máquina defeituosa. O tempo necessário para cada operação tem uma distribuição exponencial com média de 30 minutos. Entretanto, com probabilidade de $\frac{1}{3}$, a operação tem de ser realizada uma segunda vez (com o mesmo tempo de distribuição) de modo a recuperar a máquina com problema trazendo-a de volta para um estado operacional satisfatório. Um sistema de manutenção trabalha somente em uma máquina problemática por vez, realizando todas as operações (uma ou duas) necessárias a essa máquina, segundo a regra na qual os primeiros que chegam serão os primeiros a ser atendidos. Após uma máquina ser reparada, o tempo até a próxima falha tem uma distribuição exponencial com média de três horas.

**(a)** Como devem ser definidos os estados do sistema de modo a formular esse sistema de filas como uma cadeia de Markov de tempo contínuo? *Dica:* Dado que uma primeira operação esteja sendo realizada em uma máquina, ser *bem-sucedido* na finalização dessa operação e *fracassar* nessa finalização são dois eventos distintos de interesse. Use então a Propriedade 6 referente à desagregação para a distribuição exponencial.

**(b)** Construa o diagrama de taxas correspondente.

**(c)** Desenvolva as equações de equilíbrio.

**17.7-1.*** Considere o modelo *M/G/*1.

**(a)** Compare o tempo de espera previsto na fila se a distribuição de tempos de atendimento for: (i) exponencial, (ii) constante, (iii) Erlang com a parcela de variação (isto é, o desvio-padrão) a meio caminho entre os casos constante e exponencial.

**(b)** Qual é o efeito sobre o tempo de espera previsto na fila e sobre o comprimento esperado da fila caso tanto $\lambda$ quanto $\mu$ forem duplicados e a escala da distribuição de tempos de atendimento for modificada de acordo?

**17.7-2.** Considere o modelo *M/G/*1 com $\lambda = 0{,}2$ e $\mu = 0{,}25$.

T **(a)** Use o gabarito em Excel para esse modelo (ou cálculos manuais) para encontrar as principais medidas de desempenho — $L$, $L_q$, $W$, $W_q$ — para cada um dos seguintes valores de $\sigma$: 4, 3, 2, 1, 0.

**(b)** Qual é a razão entre $L_q$ com $\sigma = 4$ e $L_q$ com $\sigma = 0$? O que isso diz em relação à importância de reduzir a variabilidade dos tempos de atendimento?

**(c)** Calcule a redução em $L_q$ quando $\sigma$ é reduzido de 4 para 3, de 3 para 2, de 2 para 1 e de 1 para 0. Qual é a maior redução? Qual é a menor?

**(d)** Use o método de tentativa e erro com o gabarito para ver aproximadamente quanto $\mu$ precisaria ser aumentado com $\sigma = 4$ para atingir o mesmo $L_q$ que teria com $\mu = 0{,}25$ e $\sigma = 0$.

**17.7-3.** Considere as seguintes alternativas sobre um sistema de filas *M/G/*1, em que $\sigma^2$ é a variância dos tempos de atendimento. Classifique cada alternativa como verdadeira ou falsa e depois justifique sua resposta.

**(a)** Aumentar $\sigma^2$ (com $\lambda$ e $\mu$ fixos) aumentará $L_q$ e $L$, mas não alterará $W_q$ e $W$.

**(b)** Ao escolher entre uma tartaruga ($\mu$ e $\sigma^2$ pequenos) e uma lebre ($\mu$ e $\sigma^2$ grandes) para ser o atendente, a tartaruga sempre ganha fornecendo um $L_q$ menor.

**(c)** Com $\lambda$ e $\mu$ fixos, o valor de $L_q$ com uma distribuição exponencial de tempos de atendimento é o dobro daquele com tempos de atendimento constantes.

**(d)** Entre todas as possíveis distribuições de tempo de atendimento (com $\lambda$ e $\mu$ fixos), a distribuição exponencial resulta no maior valor de $L_q$.

**17.7-4.** Marsha opera uma banquinha de café expresso. Os clientes chegam de acordo com um processo de Poisson a uma taxa média de 30 por hora. O tempo necessário para Marsha servir um cliente tem uma distribuição exponencial com média de 75 segundos.

**(a)** Use o modelo *M/G/*1 para encontrar $L$, $L_q$, $W$ e $W_q$.

**(b)** Suponha que Marsha seja substituída por uma máquina automática de café expresso que precise exatamente de 75 segundos para cada cliente operar. Encontre $L$, $L_q$, $W$ e $W_q$.

**(c)** Qual é a razão entre $L_q$ no item (*b*) e $L_q$ no item (*a*)?

T **(d)** Use o método de tentativa e erro com o gabarito em Excel para o modelo *M/G/*1 para verificar aproximadamente quanto Marsha precisaria reduzir o tempo de atendimento esperado para alcançar o mesmo $L_q$ obtido com a máquina automática.

**17.7-5.** Antônio opera uma sapataria por conta própria. Os clientes chegam para trazer sapatos para serem consertados de acordo com um processo de Poisson a uma taxa média de 1 por hora. O tempo que Antônio precisa para reparar cada sapato (individualmente) apresenta uma distribuição exponencial com média de 15 minutos.

**(a)** Considere a formulação desse sistema de filas em que os sapatos individualmente (não o par) são considerados como os clientes. Para essa formulação, construa o diagrama de taxas e desenvolva as equações de equilíbrio, porém não as resolva.

**(b)** Considere agora a formulação deste sistema de filas no qual os pares de sapatos são considerados os clientes. Identifique o modelo de filas específico que se encaixa nessa formulação.

**(c)** Calcule o número esperado de pares de sapatos na sapataria.

**(d)** Calcule o tempo esperado do momento em que um cliente deixa um par de sapatos até eles serem consertados e estarem prontos para ser retirados pelo cliente.

T **(e)** Use o gabarito em Excel correspondente para verificar suas respostas nos itens (*c*) e (*d*).

**17.7-6.*** A base de manutenção da Friendly Skies Airline possui instalações para revisar somente um motor de avião por vez. Portanto, para que os aviões sejam liberados para uso o mais rápido possível, a política tem sido alternar a revisão dos quatro motores de cada avião. Em outras palavras, somente um motor é revisado cada vez que um avião chega no hangar. Segundo essa política, os aviões chegam de acordo com um processo de Poisson a uma taxa média de 1 por dia. O tempo necessário para uma revisão de motor (assim que os trabalhos forem iniciados) apresenta uma distribuição exponencial com média de $\frac{1}{2}$ dia.

Foi feita uma proposta para alterar a política de modo que todos os quatro motores são revisados consecutivamente cada vez que um avião chega no hangar. Embora isso quadruplicaria o tempo de atendimento esperado, cada avião precisaria ir para a base de manutenção somente um quarto das vezes.

A gerência agora precisa decidir se deve continuar na mesma situação ou adotar a nova proposta. O objetivo é minimizar o tempo de vôo médio perdido pela frota inteira por dia em razão das revisões de motor.

**(a)** Compare as duas alternativas em relação ao tempo médio de vôo perdido por um avião cada vez que ele chega na base de manutenção.

**(b)** Compare as duas alternativas em relação ao número médio de aviões perdendo tempo de vôo em virtude de se encontrar na base de manutenção.

**(c)** Qual das duas comparações seria apropriada para tomar a decisão da gerência? Explique.

**17.7-7.** Reconsidere o Problema 17.7-6. A gerência adotou a proposta, mas agora quer que sejam realizadas mais análises desse novo sistema de filas.

**(a)** Como o estado do sistema deveria ser definido de modo a formular o modelo de filas como uma cadeia de Markov de tempo contínuo?

**(b)** Construa o diagrama de taxas correspondente.

**17.7-8.** A fábrica McAllister Company possui *dois* almoxarifados, cada um deles com um *único* almoxarife, em sua área de fabricação. Um almoxarifado manipula apenas as ferramentas para o maquinário pesado; o segundo manipula todas as demais ferramentas. Entretanto, para cada almoxarifado, os ferramenteiros chegam para pegar as ferramentas a uma taxa média de 24 por hora e o tempo de atendimento esperado é de dois minutos.

Em decorrência das reclamações que os ferramenteiros que se dirigem ao almoxarifado têm de esperar muito, foi proposto que os dois almoxarifados sejam combinados de modo que qualquer um dos almoxarifes possa manipular qualquer tipo de ferramenta, conforme a demanda exija. Acredita-se que a taxa média de chegada para o almoxarifado combinado com dois almoxarifes duplicaria para 48 por hora e que o tempo de atendimento esperado continuaria a ser de dois minutos. Entretanto, não há informações disponíveis sobre a *forma* das distribuições probabilísticas para tempos entre chegadas e tempos de atendimento, de modo que não esteja claro qual seria o melhor modelo de filas a ser adotado.

Compare a situação atual com a proposta em relação ao número total esperado de ferramenteiros na(s) almoxarifado(s) e o tempo de espera previsto (incluindo atendimento) para cada ferramenteiro. Faça isso tabulando esses dados para os quatro modelos de filas considerados nas Figuras 17.6, 17.8, 17.10 e 17.11 (use $k = 2$ quando for apropriada uma distribuição de Erlang).

**17.7-9.*** Considere um sistema de filas com um único atendente com uma entrada de Poisson, tempos de atendimento com distribuição de Erlang e uma fila finita. Suponha, particularmente, que $k = 2$, a taxa média de chegada seja de 2 clientes por hora, o tempo de atendimento esperado seja de 0,25 hora e o número máximo permitido de clientes no sistema seja 2. Esse sistema pode ser formulado como uma cadeia de Markov de tempo contínuo dividindo cada tempo de atendimento em duas fases consecutivas, cada uma tendo uma distribuição exponencial com média de 0,125 hora e depois definindo o estado do sistema como $(n, p)$, em que $n$ é o número de clientes no sistema ($n = 0, 1, 2$) e $p$ indica a fase do cliente que está sendo atendido ($p = 0, 1, 2$, em que $p = 0$ significa que nenhum cliente está sendo atendido).

**(a)** Construa o diagrama de taxas correspondente. Escreva as equações de equilíbrio e depois use essas equações para encontrar a distribuição de estado estável do estado dessa cadeia de Markov.

**(b)** Use a distribuição de estado estável obtida no item (*a*) para identificar a distribuição de estado estável do número de clientes no sistema ($P_0, P_1, P_2$) e o número esperado de clientes de estado estável no sistema ($L$).

**(c)** Compare os resultados do item (*b*) com os resultados correspondentes quando a distribuição de tempos de atendimento for exponencial.

**17.7-10.** Considere o modelo $E_2/M/1$ com $\lambda = 4$ e $\mu = 5$. Esse modelo pode ser formulado como uma cadeia de Markov de tempo contínuo dividindo cada tempo entre atendimentos em duas fases consecutivas, cada uma tendo uma distribuição exponencial com média de $1/(2\lambda) = 0{,}125$ e depois definindo o estado do sistema como $(n, p)$, em que $n$ é o número de clientes no sistema ($n = 0, 1, 2, \ldots$) e $p$ indica a fase da *próxima* chegada (que ainda não se encontra no sistema) ($p = 1, 2$).

Construa o diagrama de taxas correspondente (mas não o resolva).

**17.7-11.** Uma empresa tem um técnico de manutenção para manter um grande grupo de máquinas em ordem. Tratando esse grupo como uma população solicitante infinita, as quebras individuais ocorrem de acordo com um processo de Poisson a uma taxa média de 1 por hora. Para cada quebra, a probabilidade é 0,9 de que seja necessário somente um pequeno reparo, em cujo caso o tempo necessário tem uma distribuição exponencial com média de $\frac{1}{2}$ hora. Caso contrário, seria necessário um reparo importante, em cujo caso o tempo de reparo tem uma distribuição exponencial com média de cinco horas. Como as duas distribuições *condicionais* são exponenciais, a distribuição *incondicional* (combinada) de tempos de reparo é *hiperexponencial.*

**(a)** Calcule a média e desvio-padrão dessa distribuição hiperexponencial. *Dica:* Use as relações gerais da teoria das probabilidades de que, para qualquer variável aleatória $X$ e qualquer par de eventos mutuamente exclusivos $E_1$ e $E_2$, $E(X) = E(X \mid E_1)P(E_1) + E(X \mid E_2)P(E_2)$ e $\text{var}(X) = E(X^2) - E(X)^2$. Compare esse desvio-padrão com aquele de uma distribuição exponencial com essa média.

**(b)** Quais são os $P_0$, $L_q$, $L$, $W_q$ e $W$ para esse sistema de filas?

**(c)** Qual é o valor condicional de $W$, dado que a máquina envolvida precise de um reparo importante? E para um pequeno reparo? Qual é a divisão de $L$ entre máquinas precisando dos dois tipos de reparos? *Dica:* A fórmula de Little ainda se aplica para as categorias de máquinas individuais.

**(d)** Como devem ser definidos os estados do sistema de modo a formular esse sistema de filas como uma cadeia de Markov de tempo contínuo? *Dica:* Considere que informações adicionais teriam de ser dadas, além do número de máquinas quebradas, para a distribuição condicional do tempo restante até que o próximo evento de cada tipo seja exponencial.
**(e)** Construa o diagrama de taxas correspondente.

**17.7-12.** Considere a variante de fila finita do modelo *M/G*/1, em que $K$ é o número máximo de clientes permitido no sistema. Para $n = 1, 2, \ldots$, façamos que a variável aleatória $X_n$ seja o número de clientes no sistema no instante $t_n$ quando o *enésimo* cliente acaba de ter sido atendido. Não contar o cliente que sai. Os tempos $\{t_1, t_2, \ldots\}$ são chamados *pontos de regeneração.* Além disso, $\{X_n\}$ $(n = 1, 2, \ldots)$ é uma cadeia de Markov de tempo discreto e é conhecida como uma *cadeia de Markov incorporada.* As cadeias de Markov incorporadas são úteis no estudo das propriedades de processos estocásticos de tempo contínuo como aqueles para um *modelo M/G*/1.

Considere agora o caso especial em que $K = 4$, o tempo de atendimento de clientes sucessivos seja uma constante fixa, digamos, dez minutos e a taxa média de chegada seja 1 a cada 50 minutos. Portanto, $\{X_n\}$ é uma cadeia de Markov incorporada com estados 0, 1, 2, 3. Como jamais existem mais que quatro clientes no sistema, jamais pode haver mais que três no sistema em um ponto de regeneração. Como o sistema é observado em partidas sucessivas, $X_n$ jamais pode decrescer mais que uma unidade. Além disso, as probabilidades de transições que resultam em aumentos em $X_n$ são obtidas diretamente da distribuição de Poisson.

**(a)** Encontre a matriz de transição em uma etapa para cadeia de Markov incorporada. *Dica:* Ao obter a probabilidade de transição do estado 3 para o estado 3, use uma probabilidade de 1 ou mais chegadas em vez de apenas 1 chegada e similarmente para outras transições para o estado 3.
**(b)** Use a rotina correspondente na área referente a cadeias de Markov do *Courseware* de PO para encontrar as probabilidades de estado estável para o número de clientes no sistema em pontos de regeneração.
**(c)** Calcule o número de clientes esperado no sistema em pontos de regeneração e compare-o com o valor de $L$ para o modelo *M*/*D*/1 (com $K = \infty$) na Seção 17.7.

**17.8-1.*** A Southeast Airlines é uma pequena companhia que faz pontes aéreas que atendem principalmente o estado da Flórida. O balcão de passagens em certo aeroporto tem apenas um atendente. Há duas linhas distintas — uma para passageiros de primeira classe e outra para passageiros de classe econômica. Quando o atendente está pronto para atender a mais um cliente, o próximo passageiro de primeira classe é atendido caso exista alguma fila. Caso contrário, será atendido o próximo passageiro de classe econômica. Os tempos de atendimento possuem uma distribuição exponencial com média de três minutos para ambos os tipos de clientes. Durante as 12 horas por dia em que o balcão de passagens está aberto, os passageiros chegam aleatoriamente a uma taxa média de 2 por hora para passageiros de primeira classe e 10 por hora para passageiros de classe econômica.

**(a)** Que tipo de modelo de filas se ajusta a esse sistema de filas?
T **(b)** Encontre as principais medidas de desempenho — $L$, $L_q$, $W$ e $W_q$ — tanto para passageiros de primeira classe quanto para classe econômica.
**(c)** Qual é o tempo de espera antes de o atendimento começar para clientes de primeira classe como uma fração desse tempo de espera para clientes de classe econômica?
**(d)** Determine o número médio de horas por dia em que o atendente se encontra ocupado.

T **17.8-2.** Considere o modelo com prioridades não-preemptivas apresentado na Seção 17.8. Suponha que existam duas classes de prioridades, com $\lambda_1 = 4$ e $\lambda_2 = 4$. Ao desenhar esse sistema de filas, lhe é oferecido a possibilidade de escolher entre as seguintes alternativas: (1) um atendente rápido ($\mu = 10$) e (2) dois atendentes lentos ($\mu = 5$).

Compare essas alternativas com as quatro medidas de desempenho médias usuais ($W$, $L$, $W_q$, $L_q$) para cada classe de prioridade ($W_1$, $W_2$, $L_1$, $L_2$ e assim por diante). Qual alternativa é preferível caso sua principal preocupação seja o tempo de espera previsto no *sistema* para classes de prioridade 1 ($W_1$)? Qual alternativa é preferível caso sua principal preocupação seja o tempo de espera na *fila* para classes de prioridade 1?

**17.8-3.** Considere a variante com um único atendente do modelo de prioridades não-preemptivas apresentado na Seção 17.8. Suponha que existam três classes de prioridades, com $\lambda_1 = 1$, $\lambda_2 = 1$ e $\lambda_3 = 1$. Os tempos de atendimento esperados para as classes de prioridades 1, 2 e 3 são 0,4, 0,3 e 0,2, respectivamente, e, portanto, $\mu_1 = 2{,}5$, $\mu_2 = 3\frac{1}{3}$ e $\mu_3 = 5$.

**(a)** Calcule $W_1$, $W_2$ e $W_3$.
**(b)** Repita o item (*a*) ao usar a aproximação de aplicar o modelo geral para prioridades não-preemptivas apresentado na Seção 17.8. Já que esse modelo geral supõe que o tempo de atendimento previsto é o mesmo para todas as classes de prioridades, use um tempo de atendimento esperado igual a 0,3 e, portanto, $\mu = 3\frac{1}{3}$. Compare os resultados com aqueles obtidos no item (*a*) e avalie o nível de aproximação fornecido ao se fazer essa hipótese.

T **17.8-4.*** Determinado núcleo de trabalho em uma ferramentaria pode ser representado como um sistema de filas com um único atendente, em que as tarefas chegam de acordo com um processo de Poisson, com taxa média de 8 por dia. Embora as tarefas que chegam sejam de três tipos distintos, o tempo necessário para realizar qualquer uma dessas tarefas possui a mesma distribuição exponencial, com média de 0,1 dia de trabalho. A prática tem sido a de trabalhar nas tarefas que chegam segundo a regra na qual os primeiros que chegam serão os primeiros a ser atendidos. Entretanto, é importante que tarefas do tipo 1 não esperem muito, ao passo que a espera é apenas moderadamente importante para tarefas do tipo 2 e relativamente sem importância para tarefas do tipo 3. Esses três tipos chegam com taxa média de 2, 4 e 2 por dia, respectivamente. Como todos os três tipos passaram, em média, por longos atrasos, foi proposto que as tarefas fossem escolhidas de acordo com uma disciplina de prioridades apropriada.

Compare o tempo de espera previsto (incluindo atendimento) para cada um dos três tipos de tarefas caso a disciplina da fila seja: (*a*) os primeiros que chegam serão os primeiros a ser atendidos, (*b*) prioridade não-preemptiva e (*c*) prioridade preemptiva.

T **17.8-5.** Reconsidere o problema da sala de emergências do *Hospital Municipal* conforme analisado na Seção 17.8. Suponha que as definições das três categorias de pacientes estejam ligeiramen-

te ligadas de modo a transferir casos marginais para uma categoria inferior. Conseqüentemente, somente 5% dos pacientes se qualificarão como casos críticos, 20% como casos graves e 75% como casos estáveis. Crie uma tabela mostrando os dados apresentados na Tabela 17.3 para esse problema revisado.

**17.8-6.** Reconsidere o sistema de filas descrito no Problema 17.4-6. Suponha agora que clientes tipo 1 sejam mais importantes do que clientes do tipo 2. Se a disciplina da fila fosse alterada da regra em que os primeiros que chegam serão os primeiros a ser atendidos para um sistema de prioridades com clientes do tipo 1 tendo prioridade não-preemptiva em relação a clientes do tipo 2, isso aumentaria, diminuiria ou manteria inalterado o número total esperado de clientes no sistema?

**(a)** Determine a resposta sem fazer qualquer cálculo e depois apresente o raciocínio que o levou a essa conclusão.

T **(b)** Verifique sua conclusão no item (*a*) encontrando o número total esperado de clientes no sistema sob cada uma dessas duas disciplinas de fila.

**17.8-7.** Considere o modelo de filas com a disciplina de fila de prioridade preemptiva apresentada na Seção 17.8. Suponha que $s = 1$, $N = 2$ e $(\lambda_1 + \lambda_2) < \mu$; e façamos que $P_{ij}$ seja a probabilidade de estado estável de que existem $i$ membros da prioridade de classe mais alta e $j$ membros da prioridade de classe mais baixa no sistema de filas ($i = 0, 1, 2, \ldots$; $j = 0, 1, 2, \ldots$). Use um método análogo ao apresentado na Seção 17.5 para derivar um sistema de equações lineares cuja solução simultânea é $P_{ij}$. Não obtenha realmente essa solução.

**17.9-1.** Considere um sistema de filas com dois atendentes, no qual os clientes chegam de duas origens distintas. Da origem 1, os clientes sempre chegam de 2 em 2, onde o tempo entre chegadas consecutivas de pares de clientes possui uma distribuição exponencial com média de 20 minutos. A origem 2 é, por si só, um sistema de filas com dois atendentes, que possui um processo de entrada de Poisson com taxa média de sete clientes por hora e o tempo de atendimento de cada um desses dois atendentes tem uma distribuição exponencial com média de 15 minutos. Quando um cliente completa o atendimento na origem 2, ele entra imediatamente no sistema de filas sendo considerado para outro tipo de atendimento. No último sistema de filas, a disciplina da fila é prioridade preemptiva em que clientes da origem 1 sempre têm prioridade preemptiva em relação a clientes da origem 2. Entretanto, os tempos de atendimento são independentes e distribuídos identicamente para ambos os tipos de clientes de acordo com uma distribuição exponencial com média de seis minutos.

**(a)** Concentre-se primeiramente no problema de derivar a distribuição de estado estável *somente* do número de clientes da origem 1 no sistema de filas sendo considerado. Usando uma formulação de cadeia de Markov de tempo contínuo, defina os estados e construa o diagrama de taxas para derivar de forma mais eficiente essa distribuição (mas não a derive efetivamente).

**(b)** Agora se concentre no problema de derivar a distribuição de estado estável do *número total* de clientes de ambos os tipos no sistema de filas sendo considerado. Usando uma formulação de cadeia de Markov de tempo contínuo, defina os estados e construa o diagrama de taxas para derivar de forma mais eficiente essa distribuição (mas não a derive efetivamente).

**(c)** Agora se concentre no problema de derivar a distribuição *conjunta* de estado estável do número de clientes de cada tipo no sistema de filas sendo considerado. Usando uma formulação de cadeia de Markov de tempo contínuo, defina os estados e construa o diagrama de taxas para derivar esta distribuição (mas não a derive efetivamente).

**17.9-2.** Considere um sistema de duas filas infinitas em série, em que cada uma das duas instalações de atendimento possui um único atendente. Todos os tempos de atendimento são independentes e possuem uma distribuição exponencial, com média de três minutos na instalação 1 e quatro minutos na instalação 2. A instalação 1 tem um processo de entrada de Poisson com taxa média de 10 por hora.

**(a)** Encontre a distribuição de estado estável do número de clientes na instalação 1 e depois na instalação 2. A seguir, mostre a solução em forma de produto para a distribuição *conjunta* do número nas respectivas instalações.

**(b)** Qual é a probabilidade de que ambos os atendentes se encontrem ociosos?

**(c)** Encontre o *número total* de clientes esperado no sistema e o tempo *total* de espera previsto (incluindo os tempos de atendimento) para um cliente.

**17.9-3.** Sob as hipóteses especificadas na Seção 17.9 para um sistema de filas infinitas em série, esse tipo de rede de filas, na verdade, é um caso especial de uma rede de Jackson. Demonstre que isso é verdadeiro descrevendo esse sistema como uma rede de Jackson, inclusive especificando os valores de $a_j$ e $p_{ij}$, dado $\lambda$ para esse sistema.

**17.9-4.** Considere uma rede de Jackson com três instalações de atendimento com valores de parâmetros mostrados a seguir.

| | | | | $p_{ij}$ | | |
|---|---|---|---|---|---|---|
| Instalação $j$ | $s_j$ | $\mu_j$ | $a_j$ | $i = 1$ | $i = 2$ | $i = 3$ |
| $j = 1$ | 1 | 40 | 10 | 0 | 0,3 | 0,4 |
| $j = 2$ | 1 | 50 | 15 | 0,5 | 0 | 0,5 |
| $j = 3$ | 1 | 30 | 3 | 0,3 | 0,2 | 0 |

T **(a)** Encontre a taxa de chegada total em cada uma das instalações.

**(b)** Encontre a distribuição de estado estável do número de clientes nas instalações 1, 2 e 3. Em seguida demonstre a solução em forma de produto para a distribuição conjunta do número nas respectivas instalações.

**(c)** Qual é a probabilidade de que todas as instalações tenham filas vazias (nenhum cliente aguardando para ser atendido)?

**(d)** Encontre o número total de clientes esperado no sistema.

**(e)** Determine o tempo de espera total previsto (incluindo os tempos de atendimento) para um cliente.

T **17.10-1.** Ao descrever a análise econômica do número de atendentes para ser fornecido em um sistema de filas, a Seção 17.10 introduz um modelo de custos básico no qual o objetivo é minimizar $E(\text{TC}) = C_s s + C_w L$. O propósito desse problema é permitir que você explore o efeito que os tamanhos relativos de $C_s$ e $C_w$ tem sobre o número ótimo de atendentes.

Suponha que o sistema de filas sendo considerado se encaixe no modelo *M/M/s* com $\lambda = 8$ clientes por hora e $\mu = 10$ clientes por hora. Use o gabarito em Excel do *Courseware* de PO para análise econômica com o modelo *M/M/s* para encontrar o número ótimo de atendentes para cada um dos seguintes casos.

**(a)** $C_s$ = US$ 100 e $C_w$ = US$ 10.
**(b)** $C_s$ = US$ 100 e $C_w$ = US$ 100.
**(c)** $C_s$ = US$ 10 e $C_w$ = US$ 100.

T **17.10-2.*** Jim McDonald, gerente da rede de *fast-food* McBurger, se dá conta que fornecer um atendimento rápido é a chave para o sucesso da lanchonete. Clientes que têm de esperar muito possuem grandes chances de ir para outra lanchonete da região da próxima vez. Ele estima que cada minuto a mais que um cliente tiver de esperar na fila antes de completar o atendimento lhe custe uma média de 30 centavos em perda de futuros negócios. Portanto, ele quer certificar-se de que há um número de caixas abertos suficiente para manter a fila em um mínimo. Cada caixa é operado por um empregado de meio período que passa o lanche pedido por cliente e efetua a cobrança. O custo total para cada um desses empregados é de US$ 9 por hora.

Durante o horário de almoço, os clientes chegam de acordo com um processo de Poisson em uma taxa média de 66 por hora. O tempo necessário para atender um cliente é estimado como uma distribuição exponencial com média de dois minutos.

Determine quantos caixas Jim deveria manter abertos durante o horário do almoço para minimizar o custo total esperado por hora.

T **17.10-3.** A Garrett-Tompkins Company dispõe de três máquinas copiadoras em sua sala de cópia para uso de seus empregados. Entretanto, em razão de recentes reclamações sobre tempo considerável desperdiçado esperando um copiador ficar livre, a gerência está estudando o acréscimo de uma ou mais copiadoras.

Durante as 2.000 horas de trabalho durante o ano, os empregados chegam na sala de cópias de acordo com um processo de Poisson a uma taxa média de 30 por hora. Acredita-se que o tempo que cada empregado precisa gastar em uma copiadora tenha distribuição exponencial com média de cinco minutos. Estima-se que a produtividade perdida em virtude do tempo gasto por um empregado na sala de cópia custe à empresa uma média de US$ 25 por hora. Cada copiadora é alugada por US$ 3.000 por ano.

Determine quantas copiadoras a empresa deveria ter para minimizar seu custo total esperado por hora.

## CASOS

### CASO 17.1 Reduzindo o Estoque de Itens em Fabricação

Jim Wells, vice-presidente de Manufatura da Northern Airplane Company, está exasperado. Sua visita pela unidade fabril mais importante da empresa essa manhã o deixou mau humorado. Entretanto, agora ele poderá descarregar sua raiva em Jerry Carstairs, o gerente de produção da unidade, que acaba de chegar ao escritório de Jim após sua convocação.

"Jerry, acabo de voltar de uma vistoria pela fábrica e estou muito desapontado." "Qual é o problema, Jim?" "Bem, você sabe muito bem quanto tenho enfatizado a necessidade de cortar nosso estoque de itens em fabricação." "Claro, temos dado duro nesse sentido", responde Jerry. "Bem, não tão duro quanto deveria!" Jim aumenta seu tom de voz ainda mais. "Você sabe o que descobri junto às prensas?" "Não." "Cinco chapas metálicas ainda esperando para serem conformadas em perfis de asas. E depois, logo após, na unidade de inspeção, 13 perfis de asas! O inspetor estava inspecionando uma delas, porém as outras 12 estavam ali ao lado, dormindo em berço esplêndido! Você sabe que temos algumas centenas de milhares de dólares atrelados a cada um desses perfis de asas. Portanto, entre as prensas e a unidade de inspeção, temos alguns milhões de metal terrivelmente caro simplesmente parado ali do lado. Isto não pode acontecer!"

O desgostoso Jerry Carstairs tenta responder. "Certo, Jim, estou bem ciente de que a unidade de inspeção é um gargalo. Normalmente a situação não é tão ruim quanto esta que você viu hoje pela manhã, mas sem dúvida nenhuma é um gargalo. Muito menos do que para as prensas. Você realmente nos pegou em um mau dia." "Espero que sim", replica Jim, "mas, até mesmo ocasionalmente, você precisa impedir que isso aconteça. Qual sua proposta?". Jerry agora se anima notadamente em sua resposta. "Bem, na verdade, já venho trabalhando nessa questão. Tenho uma série de propostas engatilhadas e solicitei a um analista de PO da minha equipe para analisar essas propostas fazendo um relatório com sugestões." "Ótimo", responde Jim, "fico feliz por ver que você está tentando resolver o problema. Dê a esse problema a mais alta prioridade e me informe o mais rápido possível". "Faremos isso", promete Jerry.

Eis o problema que Jerry e seu analista de PO estão resolvendo. Cada uma das dez prensas idênticas está sendo usada para conformar perfis de asas de avião a partir de grandes folhas de metal especialmente processadas. Essas chapas chegam aleatoriamente ao grupo de prensas a uma taxa média de 7 por hora. O tempo necessário para uma prensa conformar um perfil de asa de avião a partir de uma chapa de metal tem uma distribuição exponencial com média de uma hora. Quando completados, os perfis de asas chegam aleatoriamente a uma unidade de inspeção na mesma taxa média que as chapas de metal chegam às prensas (7 por hora). Um único inspetor fica em tempo integral inspecionando esses perfis de asas para certificar-se de que elas atendem às especificações. Cada inspeção leva $7\frac{1}{2}$ minutos e, portanto, ele é capaz de inspecionar oito perfis de asas por hora. Essa taxa de inspeção resul-

tou em uma quantidade média substancial de estoque de itens em fabricação na unidade de inspeção (isto é, o número médio de perfis de asas aguardando inspeção é bastante grande), além daquelas que já se encontravam no grupo de máquinas.

Estima-se que o custo desse estoque de itens em fabricação seja de US$ 8 por hora para cada chapa de metal que estiver nas prensas ou para cada perfil de asa que se encontrar na unidade de inspeção. Portanto, Jerry Carstairs fez duas propostas alternativas para reduzir o nível médio de estoque de itens em fabricação.

A Proposta 1 é usar um pouco menos de força nas prensas (o que aumentaria seus tempos médios para conformar um perfil de asa para 1,2 hora) de modo que o inspetor consiga suportar melhor sua produção. Isso também reduziria o custo de energia para operar cada máquina, de US$ 7,00 para US$ 6,50 por hora. Ao contrário, aumentar para força máxima aumentaria esses custos para US$ 7,50 por hora enquanto diminuiria o tempo médio para conformar um perfil de asa para 0,8 hora.

A Proposta 2 é substituir o atual inspetor por um inspetor mais jovem para essa tarefa. Ele é ligeiramente mais rápido (embora com alguma variabilidade em seus tempos de inspeção em virtude de sua menor experiência), de modo que ele poderia se dar melhor. Seu tempo de inspeção teria uma distribuição de Erlang com média de 7,2 minutos e um parâmetro de forma $k = 2$. Esse inspetor se encontra em uma classificação que requer um salário total (incluindo benefícios) de US$ 19 por hora, ao passo que o inspetor atual se encontra em uma classificação menor em que o salário é de US$ 17 por hora. Os tempos de inspeção para cada um desses inspetores são típicos daqueles que se encontram em uma mesma classificação.

Você é o analista de PO da equipe de Jerry Carstair ao qual foi solicitado para analisar esse problema. Ele quer que você "use as técnicas de PO mais atuais para verificar em quanto cada proposta reduziria o estoque de itens em fabricação e depois faça suas recomendações".

**(a)** Para ter uma base de comparação, comece analisando a situação atual. Determine a quantidade esperada de estoque de itens em fabricação nas prensas e na unidade de inspeção. Em seguida, calcule o custo total esperado por hora ao considerar o seguinte: o custo do estoque de itens em fabricação, o custo da energia para manter as prensas em operação e o custo do inspetor.

**(b)** Qual seria o efeito da proposta 1? Por quê? Faça comparações específicas com os resultados do item (*a*). Explique esse resultado para Jerry Carstairs.

**(c)** Determine o efeito da proposta 2. Faça comparações específicas com os resultados do item (*a*). Explique esse resultado para Jerry Carstairs.

**(d)** Dê suas sugestões para reduzir o nível médio de estoque de itens em fabricação na unidade de inspeção e no grupo de máquinas. Seja específico em suas recomendações e respalde-as com análise quantitativa como aquela realizada no item (*a*). Faça comparações específicas com os resultados do item (*a*) e cite as melhorias que suas sugestões produziriam.

## PRÉVIA DE UM CASO ADICIONAL NO CD-ROM

### CASO 17.2 Dilema das Filas

Diversos clientes furiosos estão reclamando sobre as longas esperas para conseguir contato em um *call center*. Parece que seria necessário maior número de atendentes para responder às ligações. Outra opção seria treinar os atendentes para que eles consigam responder às ligações de forma mais eficiente. Foram propostos alguns critérios possíveis para níveis de atendimento satisfatórios. A teoria das filas precisa ser aplicada para determinar como a operação do *call center* deveria ser redesenhada.