Centro de Engenharia Elétrica e Informática
Curso de Graduação em Engenharia Elétrica

SARAH JÉSSIKA DA PONTES ALBUQUERQUE

ESTUDO DA TEORIA DAS FILAS PARA DETERMINAÇÃO DE PARÂMETROS ESSENCIAIS A SEREM EXTRAÍDOS A PARTIR DO MONITORAMENTO DE UMA FILA

Campina Grande, Paraíba
Agosto de 2017

SARAH JÉSSIKA DA PONTES ALBUQUERQUE

ESTUDO DA TEORIA DAS FILAS PARA DETERMINAÇÃO DE PARÂMETROS ESSENCIAIS A SEREM EXTRAÍDOS A PARTIR DO MONITORAMENTO DE UMA FILA

Trabalho de Conclusão de Curso submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração : Processamento de Informação

Orientador: Prof. Dr. Edmar Candeia Gurjão

Campina Grande, Paraíba

Agosto de 2017

SARAH JÉSSIKA DA PONTES ALBUQUERQUE

ESTUDO DA TEORIA DAS FILAS PARA DETERMINAÇÃO DE PARÂMETROS ESSENCIAIS A SEREM EXTRAÍDOS A PARTIR DO MONITORAMENTO DE UMA FILA

Trabalho de Conclusão de Curso submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração : Processamento de Informação

Aprovado em: / /

Prof. Bruno Barbosa Albert
Avaliador

Prof. Dr. Edmar Candeia Gurjão
Orientador,UFCG

Campina Grande, Paraíba
Agosto de 2017

AGRADECIMENTOS

Começo agradecendo a minha família. Agradeço ao senhor Chico pela oportunidade de estudar em Campina Grande, uma cidade a mais de quinhentos quilômetros de onde fui criada. Agradeço a ele por estar comigo apesar da distância e agradeço por ser, além do meu pai, meu amigo. Agradeço a minha mãe, dona Tereza, por ter se esforçado tanto para me proporcionar uma boa educação e por toda preocupação que me manteve viva até a escrita desse trabalho. Agradeço ao Elias pois nas horas em que era necessário, os conselhos de irmão mais velho e as dicas de saúde tornaram o percurso mais fácil.

Aos amigos e amigas cultivados ao longo da graduação, muito obrigada pelas conversas e pelos momentos compartilhados. Em especial, aos amigos Danilo, Ana, Emanuel, Yuri, Sara, Alan, Saraiva e Eduarda por serem família sem necessidade de parentesco. Agradeço ainda aos colegas de curso que me apoiaram e vivenciaram comigo os últimos períodos da graduação.

Os mais sinceros agradecimentos ao meu companheiro de pouco tempo e muitas jornadas, Arthur. Muito obrigada por ter acompanhado este trabalho do início ao fim da forma mais próxima do que qualquer outra pessoa, fornecendo apoio de várias maneiras. Meu muito obrigada pelos ensinamentos passados, pela paciência, pela compreensão, pelo amor.

Agradeço ao professor Péricles pela oportunidade de realizar iniciação científica e ao engenheiro Thiago Euzébio pelos valiosos conhecimentos passados. Agradeço também aos professores que se dedicaram em cada aula dada durante a graduação. Agradeço ainda aos funcionários que fazem o departamento de engenharia elétrica, em especial a Adail e Tchai.

Sou grata ao Rotaract Club de Campina Grande, ao Centro Acadêmico de Engenharia Elétrica e aos voluntários e voluntárias que os formam. Nesses grupos, aprendi a pensar a comunidade em que estou inserida, aprendi que se pode optar por fazer e não fazer e que o fazer de forma eficiente é o caminho. Por fim, agradeço ao meu orientador Edmar Candeia pela oportunidade de trabalhar com o tema desse trabalho, pelo bom humor, pela orientação e por ser um exemplo como professor.

Abraços apertados aos que permitiram que essa jornada fosse melhor apreciada.

Por um mundo onde
"Hurry up and wait"
seja incomum.

RESUMO

As filas acabam se formando em diversas situações e causam desconforto tanto para quem gerencia o serviço quanto para quem o consome. Motivado pelo desconforto que as filas geram e imaginando que assim como em roteadores ou sistemas computacionais, as filas de atendimento de serviços (bancos, padarias ou supermercados), ou seja, fila de pessoas possam também ser monitoradas automaticamente, este trabalho pretende responder como o monitoramento permite analisar filas. A área que estuda os sistemas de filas é a teoria das filas. Modelar matematicamente uma fila é uma tarefa não trivial e depende da obtenção de determinados parâmetros. Para isso, foi realizado pesquisas sistemáticas com o uso da literatura da área. Encontramos que os principais parâmetros a serem monitorados são: tempos entre chegadas, duração de serviço, quantidade de servidores e a ordem de atendimento dos clientes.

Palavras-chave: teoria das filas, monitoramento de filas.

ABSTRACT

The queues end up forming in different situations and cause discomfort both for those who manage the service and for those who consume it. Motivated by the discomfort that queues generate and imagining that, just as in routers or computer systems, service queues (banks, bakeries or supermarkets), that is, queues of people can also be monitored automatically, this work intends to respond as the allows you to analyze queues. The area that studies queuing systems is queuing theory. Mathematically modeling a queue is a non-trivial task and depends on obtaining certain parameters. For this, systematic research was carried out using the literature of the area. We find that the main parameters to be monitored are: times between arrivals, length of service, number of servers and customer service order.

Keywords: Queueing theory, monitoring of queues.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Nota-se filas em supermercados, bancos, na utilização de recursos como memória ou processamento em sistemas computacionais, na entrega de pacotes em roteadores, dentre outros locais. Para todos esses casos, estudos podem ser realizados para determinar o ponto ótimo entre o custo da fila e o custo da oferta do serviço. No caso de terminais de autoatendimento em bancos, por exemplo, as filas diminuiriam se a quantidade de caixa dobrasse. No entanto, em períodos sem fila, haveria o dobro de caixas ociosos consumindo energia e espaço. A teoria das filas aparece como ferramenta para se entender e prever o comportamento de filas usando métodos probabilísticos (SZTRIK, 2016). Assim, a teoria das filas é usada como ferramenta de tomada de decisão em algoritmos de redes, em sistemas computacionais ou em estratégias de empresas para atender melhor o consumidor.

Imaginando que assim como em roteadores ou sistemas computacionais, as filas de atendimento de serviços (bancos, padarias ou supermercados), ou seja, fila de pessoas também possam ser monitoradas automaticamente. E desse monitoramento, dados como a quantidade de clientes na fila ou o tempo de espera por cliente podem ser obtidos. Tais dados permitem, através de um levantamento estatístico, caracterizar variáveis importantes tanto para analisar o desempenho quanto dimensionar um sistema de filas. As variáveis importantes, a análise de desempenho e o dimensionamento de sistemas de filas são tratadas pela Teoria das Filas (PRADO, 2009).

Um exemplo em que o monitoramento de certas variáveis em um sistema de fila aliado a Teoria das Filas fornece resultados satisfatórios é de uma cadeia holandesa de supermercados, a "*Hoogvliet*". Nestes supermercados, o número de clientes que entram em períodos de tempo determinados é contado eletronicamente. Em intervalos de tempo regulares, este número é verificado e os caixas podem ser configurados para funcionar ou não funcionar. Isso, além de permitir reduzir fortes variações na procura de serviços e equilibrar melhor as cargas de trabalho efetivas, permite que o supermercado ofertasse todos os itens do cliente gratuitamente para aqueles que não conseguirem encontrar um caixa com menos de 3 clientes (DIJK, 1997). Isso foi possível porque é pouquíssimo provável que tal situação ocorra.

Neste trabalho, é estudado como o monitoramento permite analisar filas. Para isso foi estudado a teoria das filas que permite obter modelos matemáticos a partir de certos parâmetros. Grande a parte do trabalho está dividida em dois capítulos: o de fundamentação teórica e o de monitoramento de filas. No capítulo de fundamentação teórica, a Teoria das filas é apresentada junto com uma revisão de probabilidade, estatística e processos estocásticos. No capítulo monitoramento de filas, é analisado como se dá o

monitoramento, o que monitorar e como os dados monitorados fornecem material para a análise a partir da teoria das filas.

1.1 Objetivos

A partir de uma revisão bibliográfica sobre teoria das filas, pretende-se:

- Determinar que dados devem se obtidos do monitoramento de um sistema de filas para se entender o comportamento da fila;
- Determinar um modelo matemático a partir desses dados que permita estimar variáveis do sistema de filas;
- Determinar medidas de desempenho de sistemas de filas;
- Analisar soluções para melhorar essas medidas de desempenho.

2 FUNDAMENTAÇÃO TEÓRICA

A Teoria das Filas estuda o fenômeno de congestionamento de sistemas. Ela trata de modelar sistemas de filas de forma matemática usando estatística e teoria da probabilidade. A Teoria das Filas trata ainda de fornecer formas de calcular, por exemplo, medidas de desempenho como a probabilidade de um sistema estar vazio ou o número médio de pessoas em uma fila.

Nesse capítulo, é mostrado desde conceitos básicos de probabilidade a fórmulas que permitem calcular probabilidades e medidas de desempenho dos modelos mais básicos de Teoria das Filas.

2.1 Tópicos em Estatística e Probabilidade

A pergunta inicial dessa seção é o porquê a Teoria das Filas faz uso da estatística e probabilidade? Ou ainda é por vezes considerada um braço da probabilidade? A resposta é que o comportamento de uma fila quase sempre não é determinístico, ele é aleatório. Por exemplo, não é possível prever a quantidade exata de pessoas em uma fila. É possível, no entanto, saber qual quantidade de pessoas é mais provável. Como a teoria da probabilidade tem como interesse descrever fenômenos aleatórios - portanto, sistemas de filas - ela será tratada nessa seção.

Antes de abordar de forma breve a teoria da probabilidade em si, é interessante comentar onde a estatística se encaixa na teoria das filas. Para analisar filas, dados como a quantidade de pessoas que chegam em uma fila em um espaço de tempo podem ser obtidos de um sistema. Geralmente, esses dados apresentam certa regularidade nos seus valores, conhecida como regularidade estatística (KLEINROCK, 1975). A noção dessa regularidade aparece quando experimentos aleatórios são suficientemente repetidos: por exemplo, se alguém joga um dado uma vez, é difícil prever o resultado, mas se for repetida essa experiência muitas vezes, nota-se que o número de vezes que cada resultado ocorre, dividido pelo número de jogadas, eventualmente se estabilizará em direção a um valor específico. Esse valor é a conhecido como a probabilidade de tal evento acontecer (LEON-GARCIA, 2008).

2.1.1 Espaço amostral, eventos e variáveis aleatórias

As expressões que formam o título dessa subseção são conceitos que aparecem na modelagem matemática de situações probabilísticas. Um conceito que descreve o fenômeno no qual o resultado é incerto é tratado como experimento aleatório. Uma

definição mais completa referente a um experimento aleatório é dada em (ALLEN, 1990): é um experimento cuja saída não é conhecida antecipadamente mas para qual o conjunto de todas as saídas individuais são conhecidas. Esse conjunto das saídas possíveis forma o espaço amostral S do experimento. E um subconjunto do espaço amostral S é chamado de evento.

No estudo de probabilidade, as saídas de um experimento precisam ser modeladas matematicamente e isso é feito pelo conceito de variável aleatória. De acordo com (ZUKERMAN, 2017), uma variável aleatória é uma função de valor real definida no espaço amostral. Ou seja, uma variável aleatória é uma função determinística que associa um número real a cada resultado possível de um experimento. O exemplo mais clássico é o experimento aleatório de jogar uma moeda. Os possíveis resultados para esse experimento são cara (H) ou coroa (T). O espaço amostral é $S = H, T$ e uma variável aleatória X pode atribuir $X = 1$ para quando a saída for cara e $X = 0$ para quando a saída for coroa. É interessante notar que funções que dependem da variável aleatória, também, são variáveis aleatórias.

2.1.2 Probabilidade, probabilidade condicional e independência

Como mostrado no início dessa seção, um experimento aleatório, quando suficientemente repetido, permite calcular a probabilidade dos resultados ocorrerem. Isso acontece pois a medida que o número de ensaios do experimento aumenta, espera-se que um limite seja atingido devido à noção de regularidade estatística (KLEINROCK, 1975). Daí a probabilidade ser dada pela frequência relativa - número de vezes que cada resultado ocorre dividido pelo número de realizações do experimento.

Considerando um espaço amostral S e um evento A desse espaço amostral, a probabilidade de A é a função de S e de todos os subconjuntos dele, denotada por P(A) e que satisfaz os seguintes axiomas:

1. $0 \leq P(A) \leq 1$;
2. P(S) = 1;
3. A probabilidade da união de eventos mutuamente exclusivos [1] é igual a soma das probabilidades desses eventos.

Como é comum perguntar qual a probabilidade de um evento A acontecer dado que o evento B aconteceu, existe o conceito de probabilidade condicional que modela essa

[1] Eventos mutuamente exclusivos ou disjuntos são eventos que não possuem elementos do espaço amostral em comum.

questão pela equação 2.1

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

com $P(B) \neq 0$ e sendo $P(A \cap B)$ a probabilidade conjunta de A e B. De modo similar, a equação 2.2 é a probabilidade condicional de evento B, dado o evento A.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{2.2}$$

Das equações 2.1 e 2.2, obtém-se:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{2.3}$$

Da equação da probabilidade conjunta 2.3, pode-se obter a regra de Bayes a seguir:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.4}$$

Quando um evento A ocorre e não afeta a probabilidade de um outro evento B ocorrer, é dito que os eventos A e B são independentes. As equações 2.5 e 2.6 mostram a consequência da independência entre A e B,

$$P(A|B) = P(A) \tag{2.5}$$

$$P(B|A) = P(B) \tag{2.6}$$

pois $P(A \cap B) = P(A)P(B)$.

Os eventos $A_1, A_2, A_3, ..., A_n$ são denominados mutuamente exclusivos e exaustivos quando a interseção entre eles é vazia e quando a união deles é igual ao espaço amostral S. Sendo B um evento do espaço amostral S, tem-se a equação 2.7 que calcula a probabilidade total do evento B.

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{2.7}$$

Usando as equações 2.4 e 2.7, obtém - se a equação a seguir por vezes conhecida como teorema de Bayes.

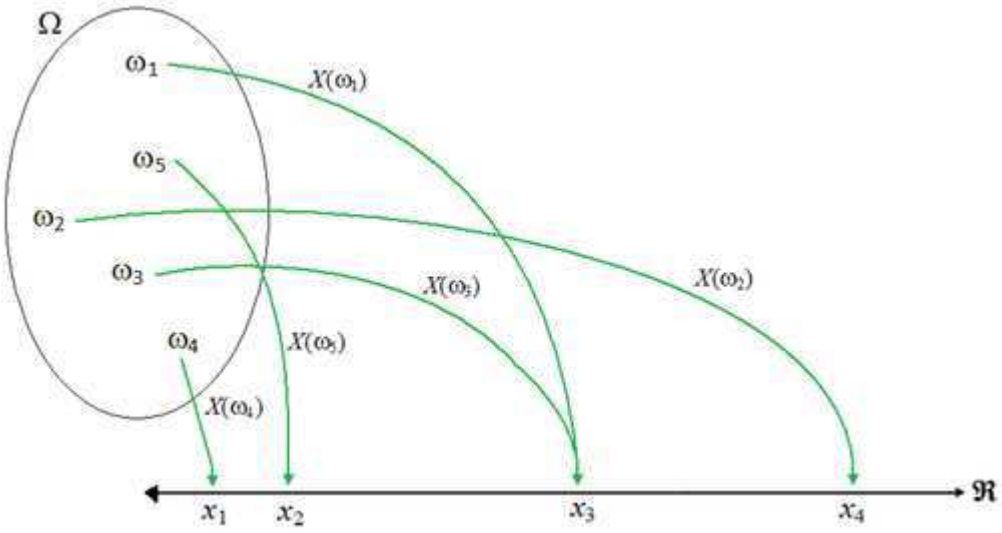
$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \tag{2.8}$$

2.1.3 Funções de probabilidade

Nesse ponto é interessante comentar a existência de variáveis aleatórias contínuas e discretas. A variável aleatória é discreta quando ela assume somente um número contável de valores distintos. Se a variável aleatória assumir um número incontável de valores possíveis, é uma variável aleatória contínua. Um exemplo usando filas de variável aleatória contínua, é o intervalo de tempo entre chegadas consecutivas. A respeito de uma variável aleatória discreta, um bom exemplo é o número de pessoas em uma fila. Ambos os tipos de variável aleatória são completamente caracterizados por funções que descrevem as medidas probabilidades de cada variável aleatória. Dessas funções, há aquelas que são específicas para variáveis aleatórias discretas e contínuas mas apresentam equivalência.

Na figura 1, é mostrado uma variável aleatória como uma função real X que atribui um único número real, denominado valor de $X(\omega)$, a cada resultado ω presente em um espaço amostral Ω . O espaço amostral Ω é o domínio e o conjunto de todos os valores de $X(\omega)$ é denominado a imagem da variável aleatória X. Percebe-se que a variável X induz

Figura 1 – Variável aleatória X como uma função real.



Fonte: https://pt.wikipedia.org/wiki/Variável_aletória.

uma medida de probabilidade na reta real como segue:

$$P(X = x) = P(\omega : X(\omega) = x) \tag{2.9}$$

$$P(X \leq x) = P(\omega : X(\omega) \leq x) \tag{2.10}$$

$$P(x_1 \leq X \leq x_2) = P(\omega : x_1 \leq X(\omega) \leq x_2) \tag{2.11}$$

Tais definições são aplicáveis para variáveis aleatória discretas e contínuas. E permitem a definição da função distribuição cumulativa de uma variável aleatória X. Tal função está definida na equação 2.12 e é conhecida também como função distribuição de probabilidade (FDP).

$$F_X(x) = P(X \leq x) \tag{2.12}$$

As propriedades de destaque da $F_X(x)$:

1. $0 \leq F_X(x) \leq 1$;
2. $F_X(x_1) \leq F_X(x_2)$ se $x_1 < x_2$;
3. $F_X(\infty) = 1$;
4. $F_X(-\infty) = 0$

Para variáveis aleatórias discretas com FDP $F_X(x)$, $F_X(x)$ é uma função escada que muda de valor apenas em saltos e é constante entre os saltos. Supondo que os saltos ocorram nos pontos $x_1, x_2, ...$ e que $x_i < x_j$ se $i < j$, tem-se a definição da função massa de probabilidade (fmp) da variável aleatória discreta X $p_X(x)$ a seguir:

$$F_X(x_i) - F_X(x_{i-1}) = P(X \leq x_i) - P(X \leq x_{i-1}) = P(X = x) \tag{2.13}$$

$$p_X(x) = P(X = x) \tag{2.14}$$

As propriedades de $p_X(x)$ são:

1. $0 \leq p_X(x_i) \leq 1$, $i = 1, 2, ...$;
2. $p_X(x) = 0$ se $x \neq x_i$;
3. $\sum_i p_X(x_i) = 1$

E é interessante perceber que a FDP $F_X(x)$ de uma variável aleatória X pode ser obtida pela equação 2.15

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p_X(x_i) \tag{2.15}$$

Para variáveis aleatória contínuas com FDP $F_X(x)$, $F_X(x)$ é contínua e também tem uma derivada $\frac{dF_X(x)}{dx}$ que existe em todos os pontos, exceto em um número finito deles,

e é contínua por partes. Daí decorre que se X for uma variável aleatória contínua, então a função massa de probabilidade de X é nula.

$$P(X = x) = 0 \tag{2.16}$$

E é definida a função de probabilidade, representada na equação 2.17. Ela é tratada através de outros termos: função densidade de probabilidade (fdp), distribuição de probabilidade ou apenas distribuição.

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{2.17}$$

As propriedades de $f_X(x)$ são as seguintes:

1. $f_X(x) \geq 0$;
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$;
3. $f_X(x)$ é contínua por partes;
4. $P(a < X \leq b) = \int_a^b f_X(x)dx$

E percebe-se que a FDP $F_X(x)$ de uma variável aleatória X pode ser obtida pela equação

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)du \tag{2.18}$$

2.1.4 Medidas Estatísticas

A esperança - média ou valor esperado - de uma variável aleatória X expressa por $E(X)$ ou μ_X é definida pela equação 2.19 para variável discreta e por 2.20 para variável aleatória contínua.

$$\mu_X = E(x) = \sum_n x_n p_X(x_n) \tag{2.19}$$

$$\mu_X = E(x) = \int_{-\infty}^{\infty} x f_X(x)dx \tag{2.20}$$

O n-ésimo momento de uma variável aleatória X é definido pela equação para o caso discreto e pela equação para o caso contínuo.

$$E(X^n) = \sum_k {x_k}^n p_X(x_k) \tag{2.21}$$

$$E(X^n) = \int_{-\infty}^{\infty} x^n f_X(x)dx \tag{2.22}$$

A variância de uma variável aleatória X expressa por ${\sigma_X}^2$ ou $Var(X)$ é definida pela equação 2.23. Ela mede a dispersão de X em relação ao valor médio $E(X)$.

$$Var(X) = {\sigma_X}^2 = E[(X - \mu_X)^2] \tag{2.23}$$

O desvio padrão de uma variável aleatória X expressa por σ_X é a medida de quanto em média os valores da variável aleatória X se desviam do valor esperado $E(X)$.

2.1.5 Distribuições de probabilidade

Nessa subseção, será discutido distribuições úteis para descrever variáveis aleatórias que modelam fenômenos de sistemas de filas.

2.1.5.1 Distribuição de Bernoulli

É uma distribuição de variável aleatória discreta e representa um experimento que tem apenas duas saídas possíveis mutuamente exclusivas. As saídas são constantemente chamadas de sucesso e falha. A variável de Bernoulli atribui o valor $X = 1$ para a saída sucesso e $X = 0$ para a saída fracasso. Sendo p a probabilidade de sucesso - como sucesso e falha são mutuamente exclusivas e exaustivas, a probabilidade de falha é $1 - p$. A função de probabilidade em termos da variável aleatória de Bernoulli é:

$$P(X = 1) = p \tag{2.24}$$

$$P(X = 0) = 1 - p \tag{2.25}$$

2.1.5.2 Distribuição Geométrica

A distribuição geométrica trata de variáveis aleatórias discretas. E é a distribuição de uma variável aleatória X que representa o número de experimentos independentes requeridos até o primeiro sucesso, cada um dos quais com p sendo a probabilidade de sucesso. Para X ser igual a n, nós devemos ter $n-1$ falhas consecutivas e então um sucesso no n experimento independente de Bernoulli . A função massa de probabilidade é dada por:

$$P(X = n) = (1 - p)^n p, n = 0, 1, 2, ... \tag{2.26}$$

Uma variável aleatória geométrica possui uma importante propriedade chamada falta de memória, *memoryless*. A variável aleatória geométrica é sem memória porque ela é baseada em experimentos independentes de Bernoulli: m falhas não afeta a probabilidade que os

próximos n experimentos resultem em falha. Uma variável aleatória é dita sem memória quando:

$$P(X > m + n|X > m) = P(X > n) \tag{2.27}$$

A distribuição geométrica é a única dentre as distribuições para variáveis discretas a apresentar a propriedade de perda de memória.

2.1.5.3 Distribuição Binomial

A distribuição Binomial trata de variáveis aleatória discretas que representam o número de sucessos em n experimentos independentes de Bernoulli. A variável aleatória binomial com parâmetros k e p apresenta a função de probabilidade dada pela equação 2.28 onde $n = 0, 1, 2, ..., k$.

$$P(X = n) = \binom{k}{n} p^n (1 - p)^{k-n} \tag{2.28}$$

A variável aleatória binomial com parâmetros 1 e p é uma variável aleatória de Bernoulli.

2.1.5.4 Distribuição de Poisson

A distribuição de Poisson trata variáveis aleatórias discretas e é aplicável para modelar o número de ocorrências de um evento em um intervalo especificado que satisfaça as seguintes condições:

- O número de ocorrência de um evento em um intervalo de tempo é independente do numero de ocorrências do evento em qualquer outro intervalo disjunto. As ocorrências apresentam independência uma da outra.
- A probabilidade de uma ou mais ocorrências simultâneas é praticamente zero.
- O número médio de ocorrências por unidade de tempo é constante ao logo do tempo, ou seja as ocorrências são distribuídas uniformemente sobre o intervalo considerado.
- O número de ocorrências durante qualquer intervalo depende somente da duração ou tamanho do intervalo; percebe-se que quanto maior o intervalo, maior o número de ocorrências.

A variável aleatória de Poisson X apresenta apenas o parâmetro λ que é interpretado como uma taxa média de ocorrência de eventos. A probabilidade de ocorrerem exatamente n eventos é dada por:

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}, n = 0, 1, 2, ... \tag{2.29}$$

É interessante saber que a soma de duas variáveis de Poisson independentes é ainda uma variável de Poisson com parâmetro igual à soma dos respectivos parâmetros. Ou seja, sendo $Y = X_1 + X_2$ onde X_i é uma variável aleatória de Poisson com parâmetro λ_i, tem - se:

$$P_Y(k) = P(X_1 + X_2 = k) = \frac{(\lambda_1 + \lambda_2)^n}{n!} e^{-(\lambda_1+\lambda_2)} \tag{2.30}$$

Outra informação interessante é que a distribuição de Poisson é um caso especial da distribuição binomial quando o número de experimentos é muito grande e a probabilidade p é pequena. Fazendo o limite da expressão Binomial quando o número de experimento tende ao infinito, obtém a distribuição de Poisson.

2.1.5.5 Distribuição Exponencial

A distribuição exponencial apresenta um forte relação com a distribuição de Poisson. Uma variável aleatória de Poisson modela o número de ocorrências em um intervalo. O tempo transcorrido entre duas ocorrências consecutivas é considerada uma variável aleatória exponencial. A distribuição exponencial, então, é usada para modelar experimentos em que a duração é variável.

A distribuição exponencial trata de variáveis aleatórias contínuas e a densidade de uma distribuição exponencial com parâmetro μ é dada por

$$f(t) = \mu e^{-\mu t}, t > 0 \tag{2.31}$$

e a função de distribuição dada por:

$$F(t) = 1 - e^{-\mu t}, t \geq 0 \tag{2.32}$$

onde μ é a taxa de ocorrência da variável aleatória. Além dessas equações, a variável aleatória exponencial é descrita, ainda, pela sua função de distribuição complementar.

$$\overline{F}(t) = e^{-\mu t}, t \geq 0 \tag{2.33}$$

A propriedade de falta de memória está presente em variáveis aleatórias distribuídas exponencialmente. Esta propriedade afirma que para todo $x \geq 0$ e $t \geq 0$, tem -se:

$$P(X > x + t | X > t) = P(X > x) \tag{2.34}$$

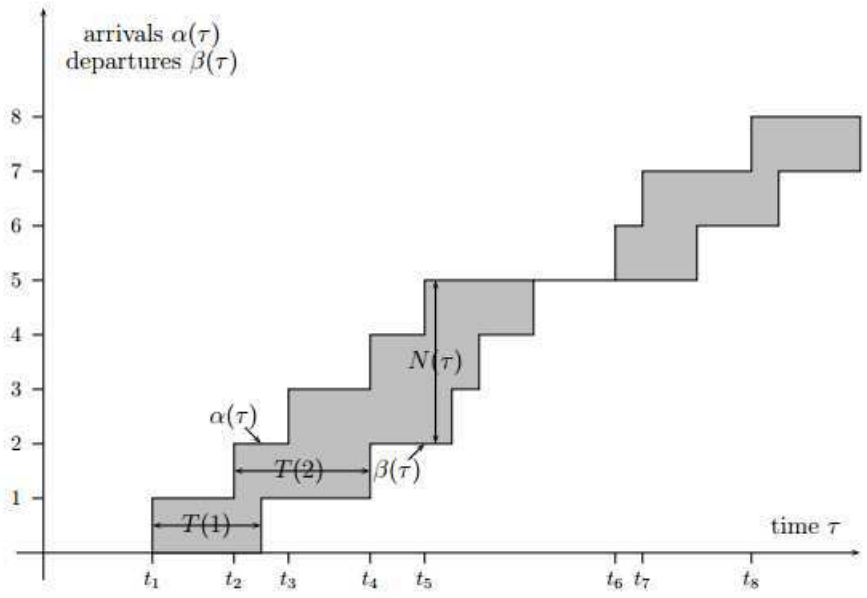
Uma interpretação da equação 2.34 é que, se X é o tempo de espera até ocorrer um evento particular e em t unidades de tempo nenhum evento ocorreu, então a distribuição de tempo de espera adicional é a mesma que seria se nenhum tempo de espera tivesse passado. O sistema, portanto, não tem a memória que as unidades de tempo não produziram o evento.

2.2 Tópicos em Processos Estocásticos

Processos estocásticos formam a parte da probabilidade que analisa a ocorrência de um fenômeno aleatório no decorrer de algum parâmetro. Os parâmetros mais comuns são o tempo e o espaço. Uma revisão sobre processos estocástico aparece neste trabalho pois as variáveis aleatórias inerentes as filas podem ser analisadas no decorrer do tempo.

Alguns exemplos de processos estocásticos que modelam sistemas de fila estão representados na figura 2 . São eles os processos $\alpha(\tau)$, $\beta(\tau)$, $T(n)$, $N(\tau)$. O processo $\alpha(\tau)$ modela a quantidade de chegadas de clientes no tempo; $\beta(\tau)$ modela a quantidade de partidas de clientes no tempo, $T(n)$ modela o tempo gasto por cliente e $N(\tau)$ modela o número de clientes no sistema no decorrer do tempo. Percebe-se que os processos estocásticos são funções que variam, em geral, no tempo onde os valores assumidos são variáveis aleatórias. Formalmente, um processo estocástico é uma família de variáveis aleatórias X_i, indexadas por um parâmetro i, onde i pertence a algum conjunto de indexadores I.

Figura 2 – Alguns processos estocásticos em uma fila.

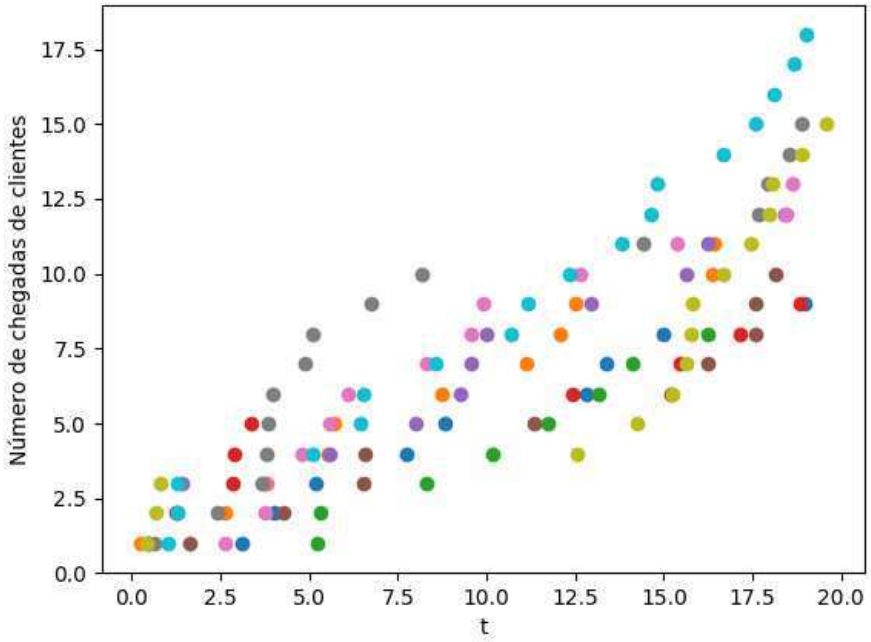


Fonte: (BERTSEKAS; GALLAGER, 1992).

Se apoiando ainda na figura 2, as variáveis aleatórias $\alpha(\tau)$ para cada tempo τ podem assumir qualquer valor que represente a quantidade de pessoas que chegaram em uma fila, $\{0, 1, 2, 3, 4, ...\}$. Esse conjunto de valores possível para a família de variáveis aleatórias $\alpha(\tau)$ é denominado espaço de estados de um processo estocástico. Os estados, portanto, são os valores assumidos por variáveis aleatória no conjunto de parâmetros indexados, neste caso o tempo τ. A figura 3 mostra alguns dos diversos resultados possíveis

para o processo $\{\alpha(\tau), \tau \in I\}$. Cada cor na figura 3 é uma simulação da chegada de clientes a uma fila. Em cada simulação, percebe-se que os valores de chegadas assumem valores diferentes em momentos diferentes. E mesmo tendo realizado cinco simulações, é impossível dizer exatamente o comportamento de uma sexta ou sétima simulação. Isso caracteriza um fenômeno aleatório e por se indexado no tempo, caracteriza um processo estocástico.

Figura 3 – Simulações do processo chegada de clientes em uma fila. Cada cor representa uma simulação.



Fonte: Própria autora.

Algumas classificações são possíveis para processos estocásticos. Dentre as classificações, há aquelas que dependem dos valores assumidos pelo espaço de estados e pelo parâmetro indexado. Esses podem assumir valores discretos ou contínuos. Daí as classificações possíveis de um processo estocástico são:

- Processo de estados contínuos e de tempo contínuo;
- Processo de estados contínuos e de tempo discreto;
- Processo de estados discretos e de tempo contínuo;
- Processo de estados discretos e de tempo discreto.

O processo $\alpha(\tau)$ é um processo de estado discreto em tempo contínuo pois os valores assumidos por α são estados discretos que variam no tempo. Os processos de estados discretos são conhecidos como cadeias.

Os processos estocásticos podem ser classificado ainda pela dependência estatística entre variáveis aleatória $X(t)$ para valores diferentes do parâmetro indexado. Para classificar dessa forma, idealmente, é necessária uma distribuição conjunta de todas as variáveis. No entanto, muitos processos interessantes permitem uma descrição mais simples (KLEINROCK, 1975). E é bastante raro, com exceção dos sistemas sem memória, que as distribuições possam ser calculadas. Geralmente, apenas as médias ou transformações podem ser calculadas (SZTRIK, 2016).

2.2.1 Tipos Interessantes de Processos Estocásticos

A seguir tipos de processos estocásticos úteis na teoria das filas são descritos:

- Processos Estacionário: um processo estocástico é dito estacionário se ele é invariante ao deslocamento no tempo, ou seja, se as distribuições não variam quando há variação no tempo.
- Processos Independentes: não há dependência entre as variáveis aleatórias. A função densidade de probabilidade conjunta é obtida pela multiplicação das funções densidades de probabilidade de cada.
- Processo de Markov: nesse processo toda história passada é resumida no estado atual, propriedade de falta de memória (*memoryless*). A propriedade de perda de memória é conhecida com propriedade de Markov.
- Processo de Nascimento e Morte: classe especial do processo de Markov onde as transições de estado são apenas entre estados vizinhos.
- Processos de Semi-Markov: classe que permite uma distribuição arbitrária do tempo gasto pelo processo em um estado. Pois se for considerado apenas os instantes de transição do estado, obtém-se uma cadeia de Markov embutida.

2.2.2 Cadeias de Markov

Uma cadeia de Markov pode representar o comportamento da quantidade de clientes em uma fila. O número de clientes em uma fila pode variar dentro de uma gama de valores, geralmente $0, 1, 2, 3, 4...k$ clientes, onde k pode ser inclusive infinito. Percebe-se que não é possível haver valores negativos ou não inteiros de clientes em uma fila. Percebe-se também que, eventualmente, o número de clientes mudará de valor. Considerando n o número de mudanças da quantidade de clientes e N_n para indicar o número de clientes

na fila, então a sequência$X_0, X_1, X_2, ..., X_n, X_{n+1}$ forma a **trajetória do processo**. Os valores que X_n podem assumir, $X_n \in 0, 1, 2, 3..., k$, representam o **estado do processo** na n-ésima variação da quantidade de clientes ou na n**-ésima transição de estados**. Como a chegada de clientes em uma fila é um fenômeno aleatório, a probabilidade de que cada estado seja ocupado após as $n + 1$ transições dada a trajetória completa ou sua história de ocupação até o tempo é dada por 2.35.

$$P[X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, ..., X_0 = x_0] \tag{2.35}$$

A dificuldade de determinar probabilisticamente o comportamento futuro de uma fila seria grande caso fosse necessário informações de todos os estados anteriores do processo como mostrado em 2.35. No entanto, como se trata de um processo de Markov, apenas o último estado da fila é relevante para determinar o comportamento futuro dela. Essa é a propriedade de Markov já apresentada neste trabalho como propriedade de perda ou falta de memória.

Um cadeia de Markov é um processo estocástico em que o estado seguinte depende apenas do estado anterior. Um conjunto de variáveis aleatórias, então, forma um processo de Markov se a probabilidade do próximo valor depende apenas do valor atual e não dos valores anteriores (KLEINROCK, 1975). Cadeias de Markov podem ser usadas para descrever sistemas que seguem uma cadeia de eventos vinculados, onde o que acontece depois depende apenas do estado atual do sistema.

Em uma cadeia de Markov, o processo se move de um estado i para o estado j com a probabilidade de transição p_{ij} dada por $P[X_{n+1} = j|X_n = i]$ onde i e j são estados discretos do sistema e $n = 0, 1, 2....$ O processo também pode permanecer no estado atual, significando $p_{ii} = P[X_{n+1} = i|X_n = i]$. Quando as probabilidades de transição não mudam para diferentes valores n, é dito que a cadeia de Markov é homogênea. Uma distribuição de probabilidade inicial, definida no espaço de estados do processo, especifica o estado inicial, $p_{i_0} = P[X_0 = i_0]$. Como o processo, após cada transição, deve ocupar um dos estados dentre os N existentes no espaço de estados, a somatória das probabilidade de transição devem ser igual a 1, nesse estado.

$$\begin{aligned} \sum_{j=1}^{N} p_{ij} &= \sum_{j=1}^{N} P(X_{m+1} = j|X_m = i) \\ &= 1 \end{aligned} \tag{2.36}$$

A equação 2.36 mostra esta propriedade matematicamente. É possível visualizar esta propriedade somando os elementos de cada linha de um matriz de probabilidades de transição P. A matriz P descreve as probabilidades de transição e é bastante utilizada no

cálculo de certas probabilidades.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix} \tag{2.37}$$

Especificada uma cadeia de Markov com as probabilidades de transição, outras probabilidades do processo podem ser calculas. A probabilidade da cadeia de Markov está no estado i_0 e ir para o estado i_n passando por uma trajetória de estados especificada é dada pela equação 2.38.

$$P[X_n = i_n, ..., X_0 = i_0] = p_{i_{n-1}}...p_{i_0 i_1} p_{i_0}(0) \tag{2.38}$$

Outro valor de interesse é a probabilidade de um determinado estado i_n ocorrer após n transições dado o estado inicial i_0. Neste caso, como a trajetória não é determinada, há várias possibilidades de trajetória com probabilidades diversas que devem ser somadas para se alcançar o resultado. A matriz de transição facilita o processo pois calculando P^n e obtendo $p_{ij}^{(n)}$ de P^n, está se obtendo a probabilidade de partir de um estado i e chegar no estado j realizando n transições.

O último resultado leva a uma importante equação para teoria das filas, a equação de Chapman - Kolmogorov, dada em 2.39. A equação é a forma em componentes da equação $P^n = P^m P^{(n-m)}$ onde $0 \leq m \leq n$. Ela expressa a lei da probabilidade total, onde a transição em n passos a partir do estado i para o estado j está condicionada ao sistema estar no estado k após m etapas.

$$p_{ij}^n = \sum_k p_{ik}^m p_{kj}^{(n-m)} \tag{2.39}$$

Outra probabilidade passível de ser calculada é a de um determinado estado ocorrer após n transições, ou seja, independente da condição inicial. Considerando $\pi_i^{(n)} = P(X_n = i)$ a probabilidade do processo está no estado i após n transições. Expressando a probabilidade de estados após n transições em um vetor de probabilidades de estados, tem-se $p_i^{(n)} = (p_{i_0}^{(n)}, p_{i_1}^{(n)}, p_{i_2}^{(n)}, ...)$. Pela lei da probabilidade total, tem-se que:

$$P\{X_1 = i\} = \sum_k P\{X_1 = i | X_0 = k\} P\{X_0 = k\} \tag{2.40}$$

É intuitivo que $\pi_i^{(1)} = \sum_k \pi_k^{(0)} p_{ki}$ ou, na forma vetorial, $\pi^{(1)} = \pi^0 P$. Estendendo esse resultado, π_n tem -se que $\pi^{(n)} = \pi^{(n-1)} P$ e pensando em recursividade, obtém -se $\pi^n = \pi^{(0)} P^n$.

2.2.2.1 Cadeia de Markov em Tempo Contínuo

As cadeia de Markov podem classificadas em termos da escala de tempo em cadeias de tempo discreto e contínuo. Até agora foi mostrado detalhes sobre cadeias de Markov de tempo discreto. No entanto, as deduções dos modelos de filas são baseados em uma classe especial da cadeia de Markov de tempo contínuo. Serão mostradas as principais diferenças entre as duas cadeias e as ferramentas necessárias da cadeia em tempo contínuo para deduções de equações em filas.

Os modelos em CTMC (*Continuous Time Markov Chains*) diferem dos modelos em DTMC (*Discrete Time Markov Chains*) basicamente por suas transições entre os estados poderem ocorrer em qualquer instante de tempo e não em pontos discretos de tempo. A taxa (CTMC) ou probabilidade (DTMC) de transição de estados do modelo ocorre obedecendo a uma lei exponencial ou geométrica respectivamente. É interessante perceber que as únicas distribuições no caso discreto e contínuo as quais gozam da propriedade de perda de memória - propriedade de Markov - são as distribuições geométricas e exponenciais (CHANIN FERNANDO L. DOTTI,).

Um modelo em cadeia de Markov, como visto, é representado por uma matriz de transição de estados. A probabilidade de cada estado em regime estacionário é a solução do sistema da equação linear:

$$\pi Q = 0 \tag{2.41}$$

onde Q é a matriz de transição de estados e π (vetor de probabilidade) é o autovetor correspondente ao autovalor unitário da matriz de transição. É importante ressaltar que a soma dos elementos do vetor de probabilidade π deve ser igual a 1.

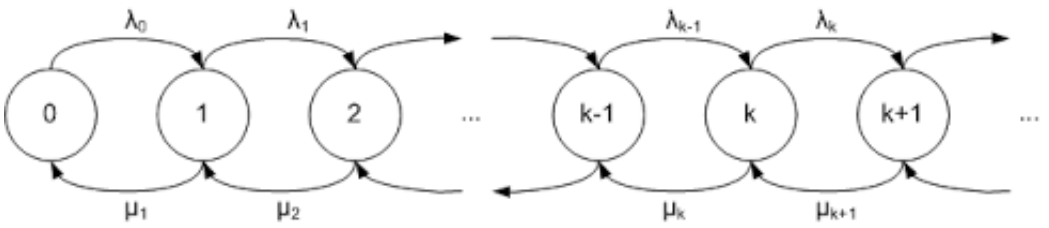
Para os modelos em CTMC, a matriz de transição de estados Q é denominada de gerador infinitesimal, onde cada elemento da linha i e coluna j da matriz , não sendo da diagonal, representa a taxa de transição do estado i para o estado j do modelo. Os elementos diagonais de Q representam o ajuste necessário para que a soma dos elementos de cada linha seja igual a zero. Para os modelos em DTMC, no entanto, a matriz de transição de estados P é denominada de matriz estocástica, onde cada elemento não sendo da diagonal representa a probabilidade de transição entre os estados do modelo. Os elementos diagonais de P representam o ajuste necessário para que a soma dos elementos de cada linha seja igual a um (CHANIN FERNANDO L. DOTTI,).

2.2.3 Processos de Nascimento e Morte

É um caso especial da cadeia de Markov em tempo contínuo onde as transições a partir do estado E_k são permitidas apenas para os estados vizinhos E_{k+1} e E_{k-1}. O processo

nascimento e morte é usado para modelar mudanças no tamanho de uma população no decorrer do tempo. Para isso é apresentado alguns conceitos. Afirmar que um processo está no estado E_k significa que a população em um tempo é do tamanho k. A expressão nascimento significa a ocorrência da transição do estado E_k para E_{k+1}. E a expressão morte significa a ocorrência da transição do estado E_k para E_{k-1}. Existem taxas que descrevem quantos nascimentos λ_k ou mortes μ_k ocorrem quando a população tem tamanho k. Essas taxas são as probabilidades de transição dadas pelas equações 2.42 e 2.43. Como tais taxas são independentes do tempo t e dependem apenas de E_k, tem-se um cadeia de Markov homogênea de tempo contínuo do tipo nascimento e morte.

Figura 4 – Diagrama de transição de estados de um processo de Nascimento e Morte



Montar a matriz de transição de um processo de nascimento e morte é interessante para consolidar os detalhes envolvidos neste tipo de processo. Sabe-se de um processo de nascimento e morte que as transições acontecem apenas entre estados vizinhos ou na permanência no estado. As probabilidade de transição p_{ij} para $i - j > 1$, então, são nulas. As probabilidade de transição entre estados vizinhos são dadas por 2.43 e 2.42 para quando há nascimento e morte respectivamente.

$$p_{k'k+1} = \lambda_k \quad para \quad k = 0, 1, ... \tag{2.42}$$

$$p_{k'k-1} = \mu_k \quad para \quad k = 1, 2... \tag{2.43}$$

Quanto as probabilidades de permanência no estado, é possível provar que são dadas por $p_{k'k} = -(\lambda_k + \mu_k)$ pois o somatório de todas os elementos de uma linha precisa ser nulo pois se trata de uma CTMC. A matriz de transição (chamada também de gerador infinitesimal em CTMC), em um processo nascimento e morte, é dada pela equação 2.44.

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & . & . & . \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & . & . & . \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & . & . & . \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix} \tag{2.44}$$

Tendo conhecimento sobre as propriedades básicas do processo, a preocupação se volta para o calcular a probabilidade do tamanho da população ser k em algum tempo dada pela equação 2.45. Isso pode ser feito usando a equação de Chapman-Kolmogorov.

$$P_{k(t)} = P_k = P[X(t) = k] \tag{2.45}$$

2.2.3.1 Dinâmicas de Chapman - Kolmogorov

Nessa subseção, será mostrado uma dedução rápida das equações de Chapman-Kolmogorov. Observando as movimentações possíveis da variável aleatória de um processo de nascimento e morte durante um intervalo $(t, t + \Delta t)$, nota-se que para o processo está no estado E_k um dos seguintes eventos ocorreu:

- nada ocorreu e a população continuou com o tamanho k;
- o tamanho da população era $k - 1$ no tempo t e ocorreu um nascimento durante o intervalo $t + \Delta t$;
- o tamanho da população era de $k + 1$ no tempo t e ocorreu uma morte durante o intervalo $t + \Delta t$.

A probabilidade de nada ocorrer é a multiplicação entre a probabilidade de estar no estado E_k, $P_K(t)$, e a probabilidade $p_{k'k}(\Delta t)$. Fazendo o mesmo raciocínio para os outros itens e considerando que as transições entre estados que não sejam vizinhos durante Δt são de ordem $o(\Delta t)$, obtém-se a seguinte equação:

$$P_k(t+\Delta t) = P_k(t)p_{k'K}(\Delta t)+P_{k-1}(t)p_{k-1'k}(\Delta t)+P_{k+1}(t)p_{k+1'k}(\Delta t)+o(\Delta t), k \geq 1 \tag{2.46}$$

Para o caso de $k = 0$, é necessário uma equação especial dada por:

$$P_o(t + \Delta t) = P_0(t)p_{00}(\Delta t) + P_1(1)p_{10}(\Delta t) + o(\Delta t) \tag{2.47}$$

Para solucionar as equações 2.46 e 2.47, é necessário as condições iniciais $P_k(0)$ para $k = 0, 1, 2, \ldots$, a equação $\sum_{k=0}^{\infty} P_k(t) = 1$ e assumir o seguinte:

$$\begin{aligned}
B_1 : P(\text{ exatamente 01 nascimento em}(t, t + \Delta t)|k) &= \lambda_k \Delta t + o(\Delta t) \\
B_2 : P(\text{ exatamente 01 morte em}(t, t + \Delta t)|k) &= \mu_k \Delta t + o(\Delta t) \\
B_3 : P(\text{ exatamente 0 nascimentos em}(t, t + \Delta t)|k) &= 1 - \lambda_k \Delta t + o(\Delta t) \\
B_4 : P(\text{ exatamente 0 mortes em}(t, t + \Delta t)|k) &= 1 - \mu_k \Delta t + o(\Delta t)
\end{aligned} \tag{2.48}$$

A partir disso, é deduzida as equações em 2.49 de Chapman-Kolmogorov que descrevem a dinâmica do sistema.

$$\frac{dP_k(t)}{dt} = -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{K+1}(t), k \geq 1 \quad (2.49)$$

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad (2.50)$$

Para mais detalhes da dedução, olhar a referência (KLEINROCK, 1975). Ainda nessa referência é mostrada outra técnica para se chegar nas equações em 2.49. A outra técnica é baseada no fato que os sistemas de fila representam um exemplo de uma classe muito mais ampla de sistemas dinâmicos, que podem ser referidos como sistemas de fluxo. Um sistema de fluxo é aquele em que algumas entidades fluem, se movem ou são transferidas através de um ou mais canais de capacidade finita para passar de um ponto para outro. Nessa outra abordagem para se chegar na equação de Chapman Kolmogorov, é dito que a taxa de mudança de fluxo dentro do estado é igual a diferença entre a taxa com que o sistema entra em E_k e a taxa que o sistema deixa E_k.

2.2.3.2 Processo de Poisson

Um processo de nascimento puro em que a taxa média de nascimento constante durante todos os estados é dito um processo de Poisson. Os parâmetros de um processo de Poisson, então, podem ser os seguintes:

$$\mu_k = 0, \quad \text{para todo } k \quad (2.51)$$

$$\lambda_k = \lambda, \quad \text{para todo } k \quad (2.52)$$

A equação de Chapman-Kolmogorov para o Processo de Poisson é dada na equação

$$\frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t), k \geq 1 \quad (2.53)$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \quad (2.54)$$

Quando a condição inicial é dada em 2.55,

$$P_k(0) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases} \quad (2.55)$$

onde o processo se inicia com nenhum membro no tempo 0, a solução para $P_0(t)$ é a equação exponencial dada em 2.56.

$$P_0(t) = e^{-\lambda t} \quad (2.56)$$

Substituindo a equação 2.56 na equação 2.53 , a probabilidade de k chegadas ocorrerem durante o intervalo de tempo t é $P_k(t)$ e é dada pela equação:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad k \geq 0, t \geq 0 \tag{2.57}$$

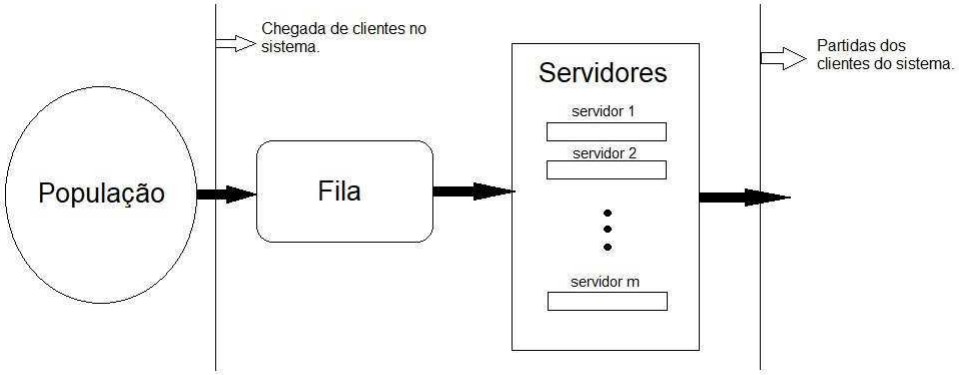
Percebe certa simplicidade nas equações de Poisson mostradas. Esse fato é interessante porque o processo de Poisson representa muitos processos reais; permitindo, portanto, obter modelos simples (KLEINROCK, 1975). Um processo real modelado constantemente pelo processo de Poisson é a chegada de cliente a uma fila. Fazendo um paralelo com a distribuição de Poisson e trazendo para a teoria das filas, percebe-se que λt representa o número médio de chegadas em um intervalo de tempo.

2.3 Teoria das Filas

2.3.1 Sistemas de Fila

Um sistema genérico de fila é representado na figura 5. Basicamente, observa-se a existência de uma população que fornece usuários ao sistema. Os usuários em espera pelo serviço constituem a fila. Os servidores prestam atendimento aos clientes e, após o atendimento, os clientes saem do sistema. Em um tratamento menos básico, é possível descrever mais ainda os sistemas de filas.

Figura 5 – Sistema de Fila Genérico



Fonte Própria

2.3.1.1 Especificação de Sistemas de Fila

Os usuários na fila são chamados de clientes. Eles chegam no sistema e, caso todos os canais de atendimentos estejam ocupados, esperam na fila para receber o atendimento. A chegadas de clientes e o tempo de atendimento se dão, geralmente, de forma aleatória.

Por isso em (TORRES, 1966), é afirmado que a formação de filas é atribuída à forma aleatória com que os clientes chegam e são atendidos pois, geralmente em sistemas reais, não é possível organizar e programar as chegadas de clientes ou os tempos de atendimento. Essas formas aleatórias são definidas como processo de chegada e processo de atendimento e são descritas estatisticamente para serem usadas constantemente em teoria das filas. Em processos que formam filas, pode haver um número máximo de clientes na fila que o sistema tolera ou pode ser considerado infinito. Por exemplo, um supermercado tem o número máximo de clientes na fila definido pela estrutura física do local. Os atendimentos aos clientes podem ser ordenados de acordo com a chegadas de clientes, o primeiro cliente a chegar é o primeiro cliente a ser atendido - por exemplo, ou pode ser de acordo com algum nível de prioridade. Há diversas formas de se ordenar o atendimento em sistemas de filas.

Percebe-se que os sistemas de filas apresentam características as quais os especificam. A primeira especificação a ser descrita é conhecida como o processo de chegada. Ela fornece informações sobre o número de clientes que solicitam serviço em um determinado tempo e é caracterizada pela distribuição do número de chegadas em intervalos disjuntos ou pelos intervalos entre tempos de chegadas sucessivas. A segunda especificação é também uma quantidade estatística e é conhecida como o processo de atendimento. Esse apresenta informações sobre a duração do atendimento. Ele é caracterizado pela distribuição do tempo de duração do serviço ou pela quantidade de pessoas atendidas em um determinado tempo. Além dessas quantidades, para especificar uma fila, é preciso se conhecer a capacidade de armazenamento disponível para armazenar clientes em espera, o tamanho da população e a disciplina de atendimento. A disciplina da fila descreve a ordem com que os clientes são atendidos.

Por existir essas características as quais especificam uma fila, utiliza-se notação de Kendall que usa letras como descritores separados por barras, A/B/m/K/N - S. A primeira posição, ocupada pela letra simbólica A, especifica o processo de chegada. A especificação acontece usando letras que indicam que tipo de distribuição de probabilidade caracteriza o processo. A segunda posição se refere ao tempo de serviço e, da mesma forma que o processo de chegada, as letras indicam a distribuição probabilidade. A terceira posição diz a quantidade de servidores no sistema. A quarta posição trata da capacidade de armazenamento do sistema. A quinta, trata do tamanho da população de clientes. A última, da disciplina na fila. Quando k e N são omitidos, são considerados infinitos.

2.3.1.2 Classificação dos Sistemas de Filas

Já tendo conhecimento do que cada parâmetro da notação A/B/m/K/N - S significa, é interessante saber que valores tais parâmetros podem assumir. As distribuições mais comuns nas quais os processos de chegada A e de atendimento B são classificados e

denominados como:

- processos de Markov, M: se refere à distribuição exponencial para os tempos do processo e à distribuição de Poisson para o numero de pessoas. É denotada por M porque, nesse tipo de processo, existe a propriedade markoviana de não ter memória, *memoryless*.
- deteminística, D: não há um processo aleatório neste caso. Os valores são constantes, ou seja, os tempos entre chegadas são todos iguais ou os tempos de atendimento são todos iguais.
- Erlang-K, E_k: distruição de Erlang com k fases, $k \geq 1$.
- Geral, G: distribuição geral sem mais especificações. Em alguns casos, a média e a variância são conhecidas.

A terceira posição assume a quantidade de servidores do sistema. A quarta posição é usada para o número de lugares disponíveis no sistema para os clientes, incluindo os espaços disponíveis nos atendimentos. Isso significa que, se houver k servidores e nenhuma sala de espera adicional estiver disponível, k será valor escrito na quarta posição. A quarta posição é omitida se a "sala de espera"for ilimitada (ZUKERMAN, 2017). A quinta posição representa o tamanho da população que pode ser infinito ou um valor constante. A sexta posição pode assumir FIFO (*First In First Out*) quando um primeiro a entrar é o primeiro a sair, ou LIFO (*Last In First Out*) ou ordem aleatória dentre diversas outras ordens. Quando essa posição é omitida, assume-se FIFO.

Exemplos de modelos de filas são os seguintes M/M/1, D/D/1, M/G/ 1, M/G/k/N. M/M/1 significa que ambos os processos de chegada e de atendimento são markovianos, que a fila tem apenas um servidor, que a capacidade do sistema é infinita e a disciplina é FIFO. D/D/1 significa que ambos os processos de chega e atendimento são determinístico, há um servido, a capacidade é infinita e a disciplina é FIFO. M/G/1 significa processo de chegada markoviano, processo de atendimento geral e apenas um atendente. M/G/k/N significa o mesmo que o anterior para as duas primeiras posições mas com k atendentes e capacidade de atendimento finita N, em que $N > k$.

Uma tendência geral na teoria da fila é a seguinte: se tanto os tempos entre chegadas quanto os tempos de serviço são distribuídos exponencialmente - são markovianos, é fácil calcular qualquer quantidade de interesse do sistema de filas. Se uma das distribuições não é markoviana, o nível de dificuldade do problema aumenta. Para o caso das filas G / G / ..., não se pode fazer muito; Mesmo os tempos médios de espera não são conhecidos. [Queueing Theory]

2.3.2 Desempenho e Eficiência

De acordo com (SZTRIK, 2016), o objetivo de todas as investigações na teoria da fila é obter medidas de desempenho do sistema que são dadas por propriedades probabilísticas (função de distribuição, função de densidade, média, variância) das seguintes variáveis aleatórias:

- Número de clientes no sistema;
- Número de clientes em espera;
- Utilização dos servidores;
- Tempo de um cliente no sistema;
- Tempo de espera de um cliente na fila;
- Tempo ocioso de um servidor;
- Tempo ocupado de um servidor.

As propriedades probabilísticas de tais variáveis, portanto as medidas de desempenho também, dependem das distribuições de tempos entre chegadas e de tempos de serviço, do número de servidores, da capacidade e da disciplina de serviço. É bastante raro, com exceção dos sistemas Markovianos, de que as distribuições possam ser calculadas. Geralmente, apenas as médias ou outros momentos podem ser calculados.

Uma vez, no entanto, especificada uma fila, ou seja, conseguindo preencher a representação A/B/m/K/M, é possível obter medidas de desempenho e eficiência do sistema. As principais medidas são:

- Número médio de clientes no sistema $\overline{N}$;
- Número médio de clientes na fila ou comprimento da fila, $\overline{N_F}$;
- Taxa de utilização dos servidores, ρ ;
- Tempo médio de um cliente no sistema, $\overline{T}$;
- Tempo médio de espera de um cliente na fila, $\overline{W}$;
- Probabilidade existirem n clientes no sistema, P_n;
- Probabilidade de o sistema está vazio, P_0.

Nesta seção, será apresentado relações referentes a taxa de utilização. O cálculo da probabilidade de existirem n clientes no sistema e da probabilidade de o sistema está vazio são mostrados apenas para os modelos markovianos na seção seguinte. Será apresentado, também, o teorema de Little pois permite relacionar as medidas de desempenho ($\overline{N}$, $\overline{N_F}$, $\overline{T}$, $\overline{W}$) de forma simples.

2.3.2.1 Taxa de Utilização

Em qualquer sistemas de filas, é possível obter as taxas médias de chegadas e atendimentos de clientes. Essa taxas são denominadas, respectivamente, em teorias das filas por λ e μ.

A taxa de utilização ou o fator de utilização ρ representa a proporção do tempo que o servidor está ocupado. É calculado pela ralação entre a taxa chegada e a taxa de atendimento dada na equação 2.58. O parâmetro c na equação diz respeito à quantidade de servidores. Os sistemas estáveis exigem que λ seja menor que μ, ou seja, $\rho < 1$. Quando ρ tende para 1 ou valores maiores, a fila tende a aumentar infinitamente. Percebe-se que é possível facilmente contornar situações em que a fila cresce infinitamente apenas aumentando a quantidade de servidores c.

$$\rho = \frac{\lambda}{c\mu} \tag{2.58}$$

2.3.2.2 Teorema de Little

O resultado de Little ou teorema de Little é uma relação fundamental entre a taxa de chegada de clientes, o número médio de clientes no sistema e o tempo médio de permanência de clientes no sistema. Representado na equação 2.59, o resultado de Little é bastante útil por conta que o processo de chegada pode ser qualquer processo estacionário e nada mais é assumido sobre o sistema.

$$\overline{N} = \lambda\overline{T} \tag{2.59}$$

Da equação 2.59, percebe-se que as medidas de desempenho número médio de cliente $\overline{N}$ e tempo de permanência no sistema $\overline{T}$ são relacionadas. No estudo da teoria da filas, geralmente, $\overline{N}$ é determinado probabilisticamente.Daí, o correspondente $\overline{T}$ é determinado usado o teorema de Little. A determinação ao contrário também pode ocorrer.

Um interpretação do teorema de Little dada por (BERTSEKAS; GALLAGER, 1992) , é se muitos clientes estão em uma fila - $\overline{N}$ é grande, então o tempo de permanência na fila será longo, $\overline{T}$ é grande. E se poucas pessoas chegam na fila - λ é pequeno, o número médio de clientes na fila é pequeno.

Na figura, percebe-se que é plotado a variação do número de clientes em uma fila no decorrer do tempo. O valor médio de $N(t)$ sobre o período de tempo t que é considerado longo pode ser calculado dividindo a área total abaixo do gráfico de $N(t)$ dividido pelo por t. Na média, cada cliente contribui com $\overline{T}$ para essa área. E o número médio de clientes chegando no intervalo t é λt. A área $\overline{N}$ é então dado por $(\lambda t)\overline{T}$.

Outra simplicidade do teorema de Little é que a equação 2.60 também é válida. Relacionando o tempo médio de permanência da fila, a taxa média de chegada e o número de pessoas na fila.

$$N_F = \lambda W \tag{2.60}$$

2.3.3 Modelos Markovianos

Os modelos Markovianos puros são denotados por M/ M/ . / . / utilizando a notação de Kendall. Isso significa que tanto o processo de chegada quanto o processo de atendimentos são markovianos. Ou seja, a distribuição do tempo entre chegadas e distribuição dos tempos de serviço são exponencialmente distribuídas.

Nesta seção, será desenvolvida uma solução geral de equilíbrio para sistemas markovianos puros. E será apresentado, a parti da solução geral, fórmulas importantes para os diversos tipos de modelos de filas markovianas.

2.3.3.1 Solução Geral de Equilíbrio

Na subseção de Processo de Nascimento e Morte, foi apresentada as equações 2.49 que permitem mostrar o comportamento transitório para um sistema de fila. O objetivo aqui, no entanto, é obter a solução de equilíbrio para essa equação, ou seja, a solução onde o sistema após um longo tempo tende a alcançar um estado estável, não mudando as distribuições de probabilidades (KLEINROCK, 1975).

A solução para as equações 2.49 é determinar $P_k(t)$. Em equilíbrio, P_k não é mais uma função no tempo. A probabilidade limite que o sistema tenha k membros em tempo arbitrário em um futuro distante é dada na equação 2.61 por p_k.

$$p_k = \lim_{t\to\infty} P_k(t) \tag{2.61}$$

Para a probabilidade de equilíbrio p_k existir, é necessário que o sistema seja ergódico e isso ocorre sempre que a sequência $\{\frac{\lambda_k}{\mu_k}\}$ permanecer abaixo 1 de alguns k em diante. Ou seja, se existir algum k_0 tal que para todo $k \geq k_0$, tem-se $\frac{\lambda_k}{\mu_k} < 1$.

Quando o sistema representado nas equações 2.49 atinge o estado de equilíbrio, o sistema não estará mais em função do tempo, resultando que as variações que ocorrem em

função do tempo são nulas, sendo $\frac{dP_k(t)}{dt}$ e $\frac{dP_0(t)}{dt}$ igualadas a zero.

$$0 = -(\lambda_k + \mu_k)p_k + \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1}, k \geq 1 \quad (2.62)$$

$$0 = -\lambda_0 p_0 + \mu_1 p_1 \quad (2.63)$$

Usando $\lambda_{-1} = \lambda_{-2} = \lambda_{-3} = 0$, $\mu_0 = \mu_{-1} = \mu_{-2} = 0$ e as probabilidades de ter um número negativo de membros da população na fila, que é nula - $p_{-1} = p_{-2} = p_{-3} = ... = 0$, consegue-se reformular a equação de Chapman-Kolmogorov de modo a usar apenas a equação 2.62. O próximo passo é determinar a solução para p_k. Substituindo $k = 1$ e calculando P_1, depois repetindo esse processo para $k = 2, 3, ...$, será notado um padrão que permite perceber que a solução para p_k é dada pela equação 2.64 a seguir.

$$p_k = \frac{\lambda_0\lambda_1...\lambda_{k-1}}{\mu_1\mu_2...\mu_k}p_0 = p_0\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}} \quad (2.64)$$

Usando a equação 2.64 e a equação $\sum_{k=0}^{\infty} p_k = 1$, obtém-se a equação que determina p_0.

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty}\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}} \quad (2.65)$$

Para todos os sistemas apresentado a seguir, as fórmulas são deduzidas a partir das equações 2.64 e 2.65 .

2.3.3.2 M/M/1

M/M/1 é considerado o sistema de fila clássico em que as taxas de nascimento e morte são constantes, $\lambda_k = \lambda$ e $\mu_k = \mu$. Nele há apenas um servidor e, como a capacidade e a disciplina da fila estão omitidas, essas são consideradas infinita e a disciplina é FIFO, *Fist In Fisrt Out*. Da equação de Chapman-Kolmogorov, consegue-se obter as equações 2.66 e 2.67 que regem o comportamento de um sistema M/M/1.

$$\frac{dp_k(t)}{dt} = -(\lambda + \mu)p_k(t) + \lambda p_{k-1}(t) + \mu p_{k+1}(t), k \geq 1 \quad (2.66)$$

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) + \mu p_1(t) \quad (2.67)$$

No entanto, na seção Solução Geral de Equilíbrio, foi deduzidas equações que permitem não usar diretamente as equações de Chapman-Kolmogorov. Como as taxas são constantes, a partir da equação 2.64, a equação 2.68 é deduzida. Ela fornece a probabilidade de k clientes estarem na fila para o sistema M/ M/ 1.

$$p_k = p_0(\frac{\lambda}{\mu})^k \quad (2.68)$$

E a equação 2.69, a partir da equação 2.65, fornece a probabilidade de não ter cliente no sistema - de o sistema está vazio ou ocioso.

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} (\frac{\lambda}{\mu})^k} \tag{2.69}$$

Sabendo a probabilidade de o sistema estar ocioso é dada por p_0, a probabilidade ou ainda proporção de tempo do servidor estar ocupado é dada por $1 - p_0$. A somatória $\sum_{k=1}^{\infty} (\frac{\lambda}{\mu})^k$ converge para $\frac{\lambda}{\mu - \lambda}$. Definindo $\rho = \frac{\lambda}{\mu}$, tem-se $\frac{\rho}{1-\rho}$. Substituindo o valor da somatória na equação 2.69, obtém-se a equação 2.70 . Por isso ρ é conhecido como o fator de utilização do sistema. E p_k é dada pela equação 2.71

$$p_0 = 1 - \rho \tag{2.70}$$

$$p_k = (1 - \rho)\rho^k \tag{2.71}$$

que representa um distribuição geométrica, compartilhando da propriedade da falta de memória. Percebe-se que quase todas distribuições de probabilidade das filas M/M/1 apresentam a propriedade de falta de memória.

Para o sistema ter uma solução de equilíbrio, é necessário que o mesmo seja ergódico. A condição necessária e suficiente para ergodicidade na fila M/ M/ 1 é que $\lambda/\mu < 1$ ou $\lambda < \mu$. A ergodicidade dá origem às probabilidades de equilíbrio p_k. Se ocorrer de $\rho > 1$, há mais chegadas do que saídas de clientes tornando o sistema instável com o número de clientes ilimitado. Se $\rho = 1$, o número de clientes que chegam, em média, é igual ao número de clientes que saem do sistema e é equiprovável qualquer quantidade de clientes no sistema, tornando o sistema instável (CHANIN FERNANDO L. DOTTI,).

Obtendo as medidas de desempenho do sistema, tem-se:

- o número médio de clientes no sistema é dado por

$$\begin{aligned} N &= \sum_{k=0}^{\infty} k p_K && (2.72)\\ &= (1 - \rho) \sum_{k=0}^{\infty} k\rho^k && (2.73)\\ &. && (2.74)\\ &. && (2.75)\\ &. && (2.76)\\ N &= \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda} && (2.77) \end{aligned}$$

- o tempo médio no sistema é dado por

$$T = \frac{N}{\lambda} \tag{2.78}$$

$$= \frac{\rho}{1-\rho} = \frac{1}{\mu - \lambda} \tag{2.79}$$

- o tempo médio na fila é dado por

$$W = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda} \tag{2.80}$$

2.3.3.3 Chegadas Desencorajadas

As chegadas desencorajadas tratam do caso em que as chegadas tendem a ser desencorajadas quando mais e mais clientes estão presente no sistema. Um forma de modelar esse efeito é escolher as taxas de natalidade e morte conforme as equações a seguir:

$$\lambda_k = \frac{\alpha}{k+1}, \quad k = 0, 1, 2, ... \tag{2.81}$$

$$\mu_k = \mu, \quad k = 1, 2, 3... \tag{2.82}$$

Aplicando a equação 2.64 e a equação 2.65 , obtém-se:

$$p_k = p_0(\frac{\alpha}{\mu})^k \frac{1}{k!} \tag{2.83}$$

$$p_0 = e^{-\alpha/\mu} \tag{2.84}$$

$$\rho = 1 - p_0 = 1 - e^{-\alpha/\mu} \tag{2.85}$$

As outras medidas de interesse são calculadas seguindo o que foi feito para o modelo M/M/1. O resultado é o seguinte:

- Número médio de clientes no sistema

$$N = \frac{\alpha}{\mu} \tag{2.86}$$

- Tempo médio no sistema

$$T = \frac{\alpha}{\mu^2(1 - e^{-\frac{\alpha}{\mu}})} \tag{2.87}$$

2.3.3.4 M/M/ ∞

O número de servidores é infinito. Isso pode ser interpretado como o caso de um servidor responsivo que acelera sua taxa de serviço linearmente quando mais clientes estão esperando ou pode ser interpretado como o caso em que sempre há um novo funcionário ou servidor disponível para cada cliente chegado (KLEINROCK, 1975). As taxas de nascimento e morte são dadas por:

$$\lambda_k = \lambda, k = 0, 1, 2... \tag{2.88}$$

$$\mu_k = k\mu, k = 0, 1, 2, ... \tag{2.89}$$

Dada as taxas, as equação 2.64 e 2.65, obtém-se

- probabilidade do sistema ter k clientes:

$$p_k = \tag{2.90}$$

- probabilidade de o sistema estar sem clientes:

$$p_0 = e^{-\frac{\lambda}{\mu}} \tag{2.91}$$

- número médio de clientes no sistema:

$$N = \frac{\lambda}{\mu} \tag{2.92}$$

- tempo médio por cliente no sistema:

$$T = \frac{1}{\mu} \tag{2.93}$$

Percebe-se que o tempo médio do cliente no sistema é o tempo médio de atendimento. Isso porque, nesse modelo, há quantos servidores for necessário.

2.3.3.5 M/M/m

O número de servidores é diferente de 1 e não é infinito. A capacidade do sistema é ilimitada. As taxas de nascimento e morte são dadas por:

$$\lambda_k = \lambda, k = 0, 1, 2... \tag{2.94}$$

$$\mu_k = k\mu, \forall\, k = 0, 1, 2, ...m,\ \mu_k = m\mu, \forall\, k > m \tag{2.95}$$

2.3.3.6 M/M/1/K

Sistema com um servidor e com capacidade de K. As taxas de nascimento e morte são dadas por:

$$\lambda_k = \lambda \forall\, k < K, \lambda_k = 0 \forall\, k \geq K \tag{2.96}$$

$$\mu_k = k\mu, k = 0, 1, 2, ..., K \tag{2.97}$$

2.3.3.7 M/M/m/m

Sistema com m servidores e com capacidade de m clientes. As taxas de nascimento e morte são dadas por:

$$\lambda_k = \lambda \,\forall\, k < m, \lambda_k = 0 \forall\, k \geq m \tag{2.98}$$

$$\mu_k = k\mu, k = 0, 1, 2..., m \tag{2.99}$$

3 MONITORAMENTO DE FILAS

O monitoramento de filas torna possível o conhecimento de certas informações que permitem tanto modelar quanto analisar o desempenho de um sistema de filas. O problema que este trabalho pretende estudar é que informações são essas. Mas especificamente, supondo o desenvolvimento de um dispositivo capaz de obter informações de sistemas de filas, que informações deveriam ser obtidas de uma fila de pessoas quando se deseja tanto estimar estados[1] quanto medir desempenho de sistemas de filas. Pretendeu-se chegar na solução dessa questão estudando a teoria das filas. Uma revisão do conteúdo estudado foi feita no capítulo anterior. Neste capítulo, pretende-se discutir que variáveis devem ser monitoradas e como elas podem ser monitoradas.

As informações de desempenho e os modelos de filas são usadas no dimensionamento de sistemas com o objetivo de prestar melhor atendimento aos clientes ou para obter redução nos custos do funcionamento do sistema. Quando tais informações são obtidas usando teoria das filas, é exigido que haja estabilidade no fluxo de chegada λ (taxa média com que os clientes chegam na fila) e na taxa de atendimentos μ (taxa média com que os clientes são atendidos), ou seja, se manterem constantes no tempo. Isso leva a restringir o tempo de observação para análise de sistemas por teoria das filas pois é preciso considerar faixas de tempos em que as taxas λ e μ são constantes. Outra exigência para estabilidade é que os atendentes sejam capazes de atender ao fluxo de chegada, isso significa que a capacidade de atendimento deve ser maior que a o ritmo de chegada de clientes$\mu > \lambda$ (PRADO, 2009).

3.1 O que monitorar?

Em teoria das filas, uma fila é identificada pelos seguinte parâmetros:

- Distribuição da chegada de clientes ou do intervalo entre chegadas;
- Distribuição do tempo de serviço;
- Número de atendentes;
- Capacidade de armazenamento do sistema;
- Tamanho da População de Clientes;
- Disciplina do atendimento.

[1] Estados se refere ao número de pessoas na fila, tempo de atendimento, tempo de espera.

Esses parâmetros, respectivamente, formam a notação de Kendall, A/B/m/K/N - S.

O dispositivo que tiver finalidade de extrair informações de uma fila deve, então, preencher a notação A/B/m/K/N - S. Os parâmetros A e B tratam de distribuições de probabilidade. Tal distribuição, quando uma grande quantidade de dados é coletada, expressa padrões do sistema. O monitoramento automático de filas permite que uma quantidade necessária de dados seja obtida para determinar distribuições de probabilidade. Essas distribuições são obtidas fazendo um análise estatísticas dos dados. Dessa análise, pode-se encontrar distribuições estatísticas que mais se assemelham ao dados reais.

3.1.1 Parâmetro A

O parâmetro A diz respeito ao processo de chegada de clientes em uma fila. Para caracterizar tal processo é necessário conhecer a distribuição da probabilidade da chegada de clientes ou a distribuição do intervalo entre chegadas. Para se chegar a tais distribuições, os tempos entre chegadas ou quantas pessoas chegaram em um determinado tempo devem ser obtidos, devem ser monitorados. No monitoramento das chegadas, é preciso contar quantas pessoas entraram na fila dentro de um intervalo de tempo determinado. No monitoramento do tempo entre chegadas, o dispositivo precisa identificar o tempo que uma pessoa entrou, em seguida, pegar o tempo que a próxima pessoa entrou no sistema e subtrair do tempo anterior.

Notaram que, quando se trata de filas, os processos de chegada geralmente seguem a distribuição de Poisson. E, para intervalos entre chegadas, a distribuição exponencial é mais representativa. Independente de ser uma distribuição de Poisson ou outra, a taxa média de chegadas λ é uma informação importante a ser calculada. Ela pode ser obtida calculando a média da quantidade de pessoas que chegam em determinado tempo ou pela média do tamanho do intervalo de tempo entre chegadas.

3.1.2 Parâmetro B

O parâmetro B diz respeito ao processo de atendimento de clientes em um sistema de filas. Para caracterizar tal processo é necessário conhecer a distribuição da probabilidade dos tempos de duração do atendimento ou quantos atendimentos são feitos em um período de tempo. Para se chegar a tais distribuições, os tempos de duração do atendimento a cada cliente ou as quantidades de clientes atendidos devem ser obtidos, devem ser monitorados.

O parâmetro B, em situações do mundo real, geralmente não seguem a distribuição exponencial nem a distribuição de Poisson. No entanto, muito dos estudos teóricos, usam tais distribuições para caracterizar o processo de atendimento. Independente da distribuição do processo de chegada, a taxa média de atendimento μ é uma informação importante a ser calculada. Ela pode ser obtida calculando a média da quantidade de pessoas que

são atendimentos em determinado tempo ou pela média do tamanho dos intervalos de atendimento.

3.1.3 Os outros parâmetros

É importante que um dispositivo que tenha interesse em obter informações de uma fila, consiga saber quantos atendentes m estão no sistema pois há dependência de tal valor na estimação de variáveis da fila. A capacidade de armazenamento do sistema é um parâmetro que geralmente não varia no tempo, podendo ser considerado uma constante e eliminado a preocupação sobre o monitoramento desse. O mesmo ocorre para o parâmetro que especifica o tamanho da população. Quanto a disciplina da fila, ela insere uma dificuldade maior na análise sendo tratada neste trabalho apenas a FIFO. Por tanto, é outro parâmetro o qual o monitoramento descrito neste trabalho não precisa se preocupar.

3.2 Análise dos dados monitorados

Usando a biblioteca Ciw, foi possível simular alguns modelos markovianos de filas. A simulação tem o objetivo de mostrar algumas das possibilidades de análise dos dados monitorados tomando como base a teoria das filas. A biblioteca Ciw é utilizada para simulação de eventos discreto para redes de filas abertas.

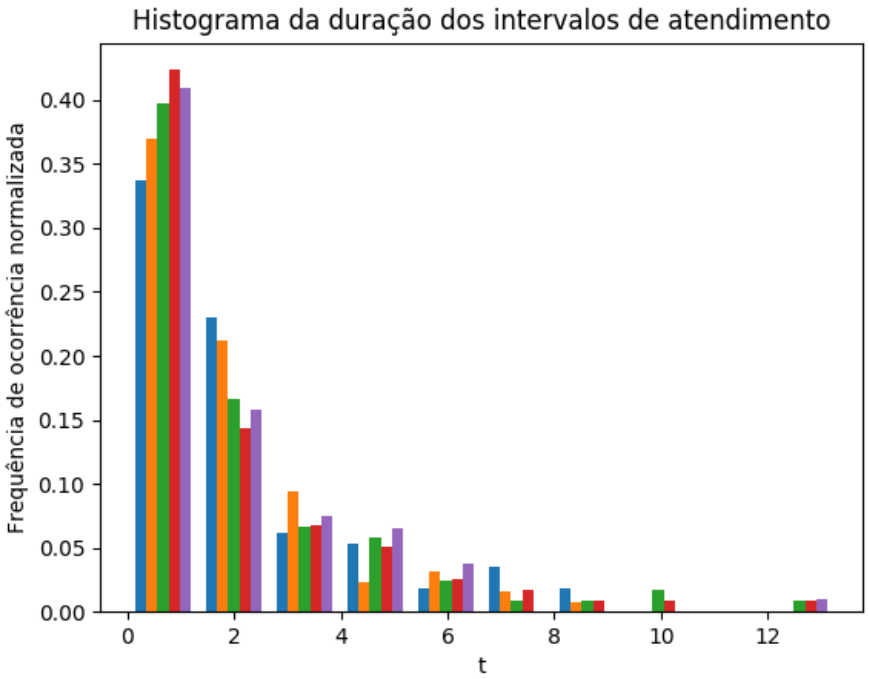
3.2.1 Processo de chegada e de atendimento

Foi simulado uma fila M/M/1 em que a taxa média de chegada λ é igual a 24 clientes por hora e a taxa de atendimento μ é igual a 30 clientes por hora. A biblioteca Ciw pede que os processos de chegada e atendimento sejam representados em termos de intervalos de tempos ao invés das taxas λ e μ. Sendo assim as distribuições são exponenciais com parâmetros 0.4 e 0.5, significando respectivamente: a cada 2.5 minutos, em média, chega um cliente e a duração dos atendimentos, em média, é 2 minutos (é mais visível tratar com esses termos do que: as taxas de chegada de clientes e de atendimentos são 0.4 clientes por minuto e 0.5 clientes por minuto respectivamente). Com essas informações, a simulação gera dados como os tempos de espera e os tempos de atendimento para vários clientes no decorrer de quatro horas, que foi o tempo da simulação.

Quando se obtém amostras de dados simulando apenas uma vez, é muito provável que tal amostra não seja confiável. O mais sensato é realizar a simulação várias vezes. O mesmo deve ocorrer para obtenção de dados em sistemas reais, realizar um número suficiente de experimento. Em (PRADO, 2009), para que os dados sejam confiáveis, a escolha de um tamanho correto de amostra é fundamental. Serão, portanto, feitas várias simulações pois assim se espera obter resultados mais confiáveis.

Imaginando que os dados gerados não foram fruto de uma simulação mas de um sistema real, o primeiro passo para analisar os dados é tentar encontrar as distribuições de probabilidade do processo de chegada e de atendimento. Isso pode ser feito gerando o histograma dos dados como feito para os tempos de atendimento e os tempos entre chegadas mostrados nas figuras[2] 6 e 7. Os dados dos histogramas foram obtidos a partir de cinco simulações; as cores representam o conjunto de dados de cada simulação. A pergunta, após feito o histograma dos dados, é que distribuição se assemelha mais com o comportamento mostrado nas figuras 6 e 7 ?

Figura 6 – Histograma dos dados de tempo de atendimento.

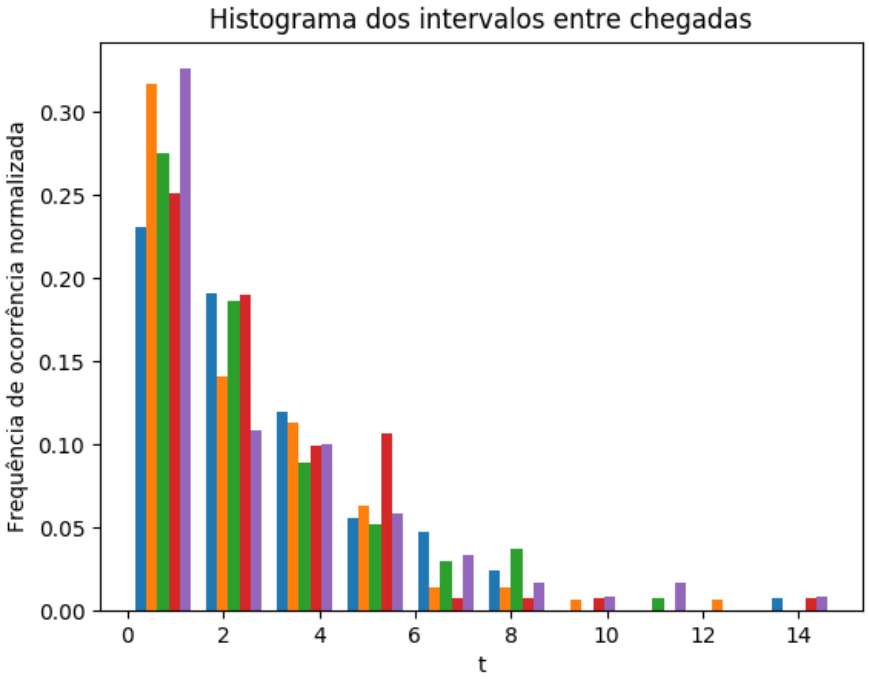


Fonte: Próprio autor.

Na simulação foi usado os modelos markovianos para as chegadas e atendimentos, então é esperado que o histograma mostre um comportamento exponencial para os processos. É o que acontece conforme as figuras mostram. Nesse ponto, é interessante comentar sobre o uso do modelo exponencial tanto no processo de chegada quanto no atendimento. Para situações reais, os intervalos de tempo entre chegadas geralmente são bem representados pelo modelo exponencial. Para os intervalos de tempo de atendimento, tal afirmação não é verdadeira para a maioria dos casos. Em situações reais, a maioria dos intervalos de tempo de atendimento não tomam os menores valores possíveis como as é representado em distribuições exponenciais.

[2] Os histogramas foram obtidos usando dados obtidos apartir de cinco simulações.

Figura 7 – Histograma dos dados de intervalos de tempos entre chegadas.



Fonte: Próprio autor.

3.2.2 Obtendo medidas de desempenho

Ainda usando a simulação da fila M/M/1 anterior, serão obtidas as medidas de desempenho do sistema de fila simulado.

A biblioteca Ciw permite obter os tempos de serviço e os tempos entre chegadas de clientes. Calculando a média desses tempos, consegue se obter o tempo de atendimento médio ($\overline{Ta}$) e o intervalo médio entre chegadas ($\overline{I}$). Esses valores determinam λ e μ usando as seguintes fórmulas:

$$\lambda = 1/\overline{I} \tag{3.1}$$

$$\mu = 1/\overline{Ta} \tag{3.2}$$

Os valores para λ e μ para realizar a simulação foram respectivamente 0.4 e 0.5. Com os dados obtidos da simulação, deve ser possível calcular λ e μ. Para a realização de 1 simulação foram obtidos os valores $\lambda = 0.361$ e $\mu = 0.476$. Para a realização de 5 simulações, foram obtidos os valores $\lambda = 0.381$ e $\mu = 0.503$. Para a realização de 50 simulações, foram obtidos os valores $\lambda = 0.412$ e $\mu = 0.498$. Observando que quanto mais simulações, mais o valores convergem para os usados na simulação.

A partir dos valores de λ e μ, consegue-se calcular as medidas de desempenho usando

as fórmulas apresentadas no capítulo de Fundamentação Teórica. Os valores encontrados, usando λ e μ encontrado a partir de 5 simulações foram os seguintes:

$$\begin{aligned} \overline{N} &= 3.12747817625881 \\ \overline{T} &= 8.205330665139792 \\ p_0 &= 0.24227868865594115 \\ \rho &= 0.7577213113440588 \end{aligned} \tag{3.3}$$

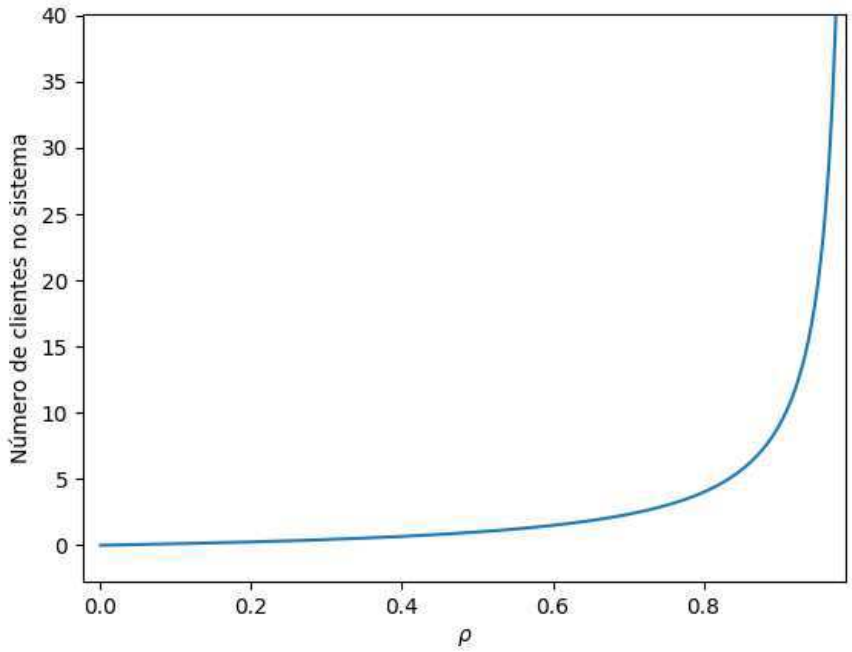
Ou seja, há três clientes no sistema em média, o tempo de permanência no sistema é 8,2 minutos em média, a probabilidade de o sistema está vazio é de 24% e a taxa de utilização do sistema é de 75%.

3.2.3 Analisando o desempenho

Obtendo-se os parâmetros de desempenho de sistemas de filas, decisões podem ser tomadas para melhorá-los. Em geral, tenta-se modificar a taxa de utilização ρ variando o número de atendentes.

Um gráfico mostrando a relação entre $\overline{N}$ e ρ é mostrado na figura 8 . Percebe-se que

Figura 8 – Variação do número médio de clientes em um sistema em função da taxa de utilização.



Fonte: Próprio autor.

quanto mais próxima de um a taxa de utilização for, maior será a quantidade de clientes no sistema. Isso implica a depender de quão próximo a taxa esteja do valor unitário, que os atendentes não serão capazes de atender ao fluxo de chegada. Portanto, quando se quer diminuir o número médio de clientes no sistema, é necessário diminuir a taxa de utilização. Para tal, pode se aumentar as taxas de atendimento ou aumentar o número de atendentes.

4 CONCLUSÃO

O curso de Engenharia Elétrica na Universidade Federal de Campina Grande conta com o tópico introdução à teoria das filas em uma das disciplinas obrigatórias, a disciplina processos estocásticos. Em outras disciplinas, como em redes de computados e arquiteturas avançadas, o assunto também foi pincelado pois aparece o roteamento de pacotes e gerenciamento de recursos respectivamente. No entanto, nessas disciplinas, apenas uma conceituação mais matemática ou uma apresentação em termos de gargalos (onde congestiona) é feita. As disciplinas de fato essenciais para se estudar filas foram as disciplinas de Probabilidade e Estatística e a de Processos Estocásticos.

Neste trabalho de conclusão de curso, foi feita uma revisão sobre teoria filas levando em consideração a teoria da probabilidade, estatística e processos estocástico. Em seguida, é dada uma visão de como se daria a aplicação da teoria das filas quando se monitora sistemas que congestionam. Os estudos levaram a entender as filas como fenômenos inevitáveis que acontecem em diversas situações; sendo por este motivo que diversas áreas fazem uso desta teoria. Levaram, também, a perceber que medidas de desempenho, usando teoria das filas, podem ser determinadas quando se obtém dados de sistemas de filas a partir de um monitoramento desses. Os dados permitem determinar parâmetros que especificam um modelo de fila o qual tem fórmulas que calculam os parâmetros de desempenho.

Neste trabalho, apenas os modelos markovianos puros foram apresentados. Com sugestão de trabalho futuros, mais modelos deveriam ser adicionados ao estudo pois filas de serviço geralmente não conseguem ser bem representadas pelos modelo M/M/./.. Outra sugestão, é realizar uma pesquisa sobre o que já está feito em termos de monitoramento e simulação de filas. A última sugestão é um estudo estatístico para entender como determinar as distribuições de tempos de chegadas e tempos de atendimentos baseados em dados reais obtidos.

REFERÊNCIAS

ALLEN, A. O. *Probability, Statistics, and Queueing Theory with Computer Science Applications*. San Diego, CA, USA: Academic Press Professional, Inc., 1990. ISBN 0-12-051051-0. Citado na página 14.

BERTSEKAS, D.; GALLAGER, R. *Data Networks (2Nd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1992. ISBN 0-13-200916-1. Citado 2 vezes nas páginas 22 e 35.

CHANIN FERNANDO L. DOTTI, P. F. A. S. R. *Avaliação Quantitativa de Sistemas*. Acessado em 10 de agosto de 2017. Disponível em: <http://www.inf.pucrs.br/~paulof/ads/AQS-tudo.pdf>. Citado 2 vezes nas páginas 27 e 38.

DIJK, N. M. van. Why queuing never vanishes. *European Journal of Operational Research*, v. 99, n. 2, p. 463 – 476, 1997. ISSN 0377-2217. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0377221797000155>. Citado na página 11.

KLEINROCK, L. *Theory*. [S.l.]: Wiley-Interscience, 1975. v. 1. ISBN 0471491101. Citado 8 vezes nas páginas 13, 14, 24, 25, 30, 31, 36 e 40.

LEON-GARCIA, A. *Probability, Statistics, and Random Processes for Electrical Engineering*. Third. Upper Saddle River, NJ: Pearson/Prentice Hall, 2008. ISBN 9780131471221 0131471228. Citado na página 13.

PRADO, D. *Teoria das Filas e Simulação*. 4. ed. Nova Lima, Minas Gerais: INDG, 2009. Citado 3 vezes nas páginas 11, 42 e 44.

SZTRIK, J. *Basic Queuing Theory*. [S.l.]: University of Debrecen, Faculty of Informatics, 2016. ISBN 9783639734713. Citado 3 vezes nas páginas 11, 24 e 34.

TORRES, O. F. Elementos da teoria das filas. *Revista de Administração de Empresas*, scielo, v. 6, 1966. ISSN 0034-7590. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-75901966000300005&nrm=iso>. Citado na página 32.

ZUKERMAN, M. *Introduction to Queueing Theory and Stochastic Teletraffic Models*. 2017. Acessado em 10 de junho de 2017. Disponível em: <https://arxiv.org/pdf/1307.2968>. Citado 2 vezes nas páginas 14 e 33.

Anexos