Resumo de Modelagem e Avaliação de Desempenho

1. Definições

Processo estocástico: é um conjunto de variáveis aleatórias $X(t), t \in T$ que descreve a evolução de um sistema ao longo do tempo sob influência do acaso
Cada $X(t)$ representa o estado aleatório do sistema no instante t

Espaço Amostral (Ω): Conjunto de todas as saídas possíveis de um experimento aleatório.

Evento: é um subconjunto qualquer de Ω

Probabilidade Simétrica: é assumir que as saídas possíveis (Experimento ou Ω) são equiprováveis (tem as mesma probabilidade)

Probabilidade Frequencista: a probabilidade $P(E)$ de um evento E é dado pela razão entre o nº de resultados favoráveis e o nº total de resultados

$$P(E) = \lim_{n \to \infty} \frac{\text{Nº de ocorrências E}}{n}$$

Probabilidade Condicional: Para eventos A e B, a probabilidade condicional de *A dado B* é definida como

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

Probabilidade Conjunta (Dependência):

$$(A \cap B) = (A|B) \cdot (B)$$

Independência: Dois eventos A e B são independentes se

$$(A \cap B) = (A)(B)$$

Teorema Probabilidade Total: afirma que, se um conjunto de eventos $(B_1, B_2, \ , B_n)$ forma uma **partição** do espaço amostral (isto é, são mutuamente exclusivos e cobrem todo o espaço), então a probabilidade de um evento (A) pode ser expressa como:

$$P(A) = \sum_{i=1}^{n} P(A \ \ B_i)\, P(B_i)$$

Regra de Bayes: Se A e B são eventos com probabilidade positiva, então

$$P(A|B) = \frac{P(B \ \ A) \cdot P(A)}{P(B)} \quad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

obs: $P(A \cap B) = P(A) \cdot (B|A)$
$P(A|B) + P(A|B^c) = 1$

Eventos mutuamente exclusivos: $P(A \cup B) = P(A) + P(B)$

Complemento:

- $P(A^c) = 1 - P(A)$
- $(A \cap B)^c = A^c \cup B^c$
- $(A \cup B)^c = A^c \cap B^c$

Função de Massa de Probabilidade (PMF)

Aplica-se a **variáveis aleatórias discretas**.
Ela fornece a probabilidade de a variável assumir um valor específico:

$$P(X = x) = f(x)$$

Deve satisfazer:

$$0 \leq f(x) \leq 1 \quad \text{e} \quad \sum_x f(x) = 1$$

Função de Densidade de Probabilidade (PDF)

Aplica-se a **variáveis aleatórias contínuas**.
Ela descreve a **densidade de probabilidade**, e não a probabilidade direta.
A probabilidade de X estar em um intervalo $[a, b]$ é:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Deve satisfazer:

$$f(x) \geq 0 \quad \text{e} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

Linearidade da Esperança

A **linearidade da esperança** afirma que a **esperança (ou valor esperado)** de uma soma de variáveis aleatórias é igual à soma das esperanças individuais, **independentemente de haver dependência entre elas**:

$$E[aX + bY] = aE[X] + bE[Y]$$

ou, mais geralmente,

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

2. Variáveis Aleatórias Discretas

Modelo Bernoulli

> Sucesso ou Fracasso

$$X \sim Ber(p) \qquad (0 < p < 1)$$

$$\mathbb{P}_X(x) = p^x(1-p)^{1-x} \quad \text{para } x \in \{0,1\}$$

- $E[X] = P$
- $Var(X) = P(1-P)$

$$\text{PMF:} \qquad\qquad\qquad \text{CDF:}$$

$$P(X=x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0 \\ 1-p, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$$

Modelo Binomial

$X \sim Bin(n,p)$

> Chama-se de experimento binomial ao experimento que
>
> - consiste em n ensaios de Bernoulli
> - cujo ensaios são independentes, e
> - para qual a probabilidade de sucessos em casa ensaio é sempre igual a p $(0 < p < 1)$

- PMF

$$p_X = (X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x},$$
$$\binom{n}{x} = \frac{n}{x(n-x)} \qquad\qquad x \in \{0, 1, 2, \ \ , n\}$$

- CDF

$$F(k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} = \sum_{y \le k}^{k} p_x^{(y)}$$

- $E[X] = n \cdot p$
- $Var(X) = n \cdot p \cdot (1-p)$

Modelo Geométrico

$X \sim eom(p)$

> Número de repetições de um ensaio de Bernoulli com probabilidade de sucesso $(0 < p < 1)$até ocorrer o primeiro sucesso

- PMF:

$$(X = x) = p \cdot (1-p)^x \qquad , x \in$$

- CDF:

$$F(k) = 1 - (1-p)^k$$

- $E[X] = \frac{1}{p}$
- $Var(X) = \frac{1-P}{p^2}$

Modelo Poisson

$X \sim Poi()$

> Nº de eventos que ocorrem em um intervalo de tempo ou espaço

- PMF:

$$p(x) = e^{-} \cdot \frac{^x}{x}, \quad x = \{0, 1, 2, \ \}$$

- CDF:

$$F(k) = \sum_{i=0}^{k} p(x)$$

$E[X] = Var(x) =$

Exponencial

- **Descrição:** Tempo até o primeiro evento (processo de Poisson).
- **PDF:**

$$f(x) = e^{-x}, \quad x \geq 0$$

- **CDF:**

$$F(x) = 1 - e^{-x}$$

Uniforme Contínua

> **Descrição:** Todos os valores em $[a, b]$ igualmente prováveis.

- **PDF:**

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{caso contrário} \end{cases}$$

- **CDF:**

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases}$$

Normal (Gaussiana)

> **Descrição:** Distribuição simétrica em torno da média .

- **PDF:**

$$f(x) = \frac{1}{\sqrt{2\pi^2}} e^{-\frac{(x-)^2}{2^2}}$$

- **CDF:**

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi^2}} e^{-\frac{(t-)^2}{2^2}}\, dt$$

(sem forma fechada; usa tabelas ou funções computacionais).

3. Relação Poisson Exponencial (Processo de Poisson)

A **distribuição de Poisson** e a **distribuição Exponencial** estão intimamente ligadas — elas descrevem **dois lados do mesmo processo estocástico**, o **Processo de Poisson**.

Processo de Poisson

Modela o número de ocorrências de um evento em um intervalo de tempo:

$$P(N(t) = k) = \frac{e^{-t}(t)^k}{k}$$

onde ( ) é a **taxa média de eventos por unidade de tempo**.

Assim, ($N(t)$) segue **distribuição de Poisson**.

Tempo entre eventos → Exponencial

O **tempo entre dois eventos consecutivos**, chamado de *tempo de interchegada*, segue uma **distribuição Exponencial**:

$$f_T(t) = e^{-t}, \quad t \geq 0$$

com média ($E[T] = 1$).

Portanto:

- Poisson → **quantos eventos ocorrem em um tempo fixo**.
- Exponencial → **quanto tempo até o próximo evento**.

Relação formal

Se os tempos entre eventos ($T_1, T_2,$) são independentes e Exponenciais ( ),
então o número total de eventos até o tempo (t):

$$N(t) = \max n : T_1 + T_2 + \ + T_n \leq t$$

segue uma **distribuição de Poisson(t)**.

E reciprocamente, se ($N(t)$) é um processo de Poisson, então os tempos entre eventos são **Exponenciais( )**.

Aspecto	Distribuição	Interpretação
Número de eventos em tempo fixo	**Poisson(λt)**	Contagem de ocorrências
Tempo entre eventos	**Exponencial(λ)**	Intervalo entre ocorrências

4. Geração de Amostras Aleatórias

Motivo: Poder simular/observar fenômenos aleatórios

Premissa: Temos um gerador de números uniformemente distribuídos entre 0 e 1: $]0, 1[$

Métodos principais

Método da Transformada Inversa

- **Ideia:** usar a função de distribuição acumulada (CDF) (F(x)) da variável desejada.
- **Passos:**
  (1) Gere ($U \sim \text{niorme}(0, 1)$);
  (2) Calcule ($X = F^{-1}(U)$).
- **Justificativa:** se (U) é uniforme em $[0, 1]$, então ($X = F^{-1}(U)$) tem CDF ($F(x)$).
- **Vantagens:** simples, exato.
- **Limitações:** exige que (F^{-1}) tenha forma analítica fácil.
- **Exemplo:**
  Exponencial( ): ($X = -\frac{1}{}\ln(1 - U)$).

Método da Aceitação–Rejeição

- **Ideia:** gerar amostras de uma distribuição difícil usando outra mais simples.
- **Passos:**
  (1) Escolha uma distribuição fácil ($g(x)$) e uma constante (c) tal que ($f(x) \leq c, g(x)$) para todo (x);
  (2) Gere ($X \sim g(x)$) e ($U \sim U(0,1)$);
  (3) Aceite (X) se ($U \leq \frac{f(X)}{c,g(X)}$), senão rejeite e repita.
- **Vantagens:** útil quando (F^{-1}) é complexa.
- **Limitações:** pode ser ineficiente se (c) for grande (muitas rejeições).

Método do Vetor (ou Método de Composição)

- **Ideia:** gerar amostras quando a distribuição é composta ou mistura de várias partes.
- **Passos:**
  (1) Escolha qual componente gerar (segundo probabilidades associadas);
  (2) Gere a amostra da distribuição correspondente.
- **Exemplo:**
  Se (X) vem de uma mistura de duas exponenciais:
  $f(x) = pf_1(x) + (1-p)f_2(x)$
  Então:
  (1) Gere ($U \sim U(0,1)$);
  (2) Se ($U < p$), gere ($X \sim f_1$); caso contrário, ($X \sim f_2$).
- **Aplicação:** simulação de **sistemas com múltiplos regimes** ou **processos compostos**.

Método	Quando usar	Exemplo típico
Transformada Inversa	CDF invertível	Exponencial, Uniforme
Aceitação–Rejeição	CDF complexa	Normal, Gamma
Vetor (Composição)	Mistura de distribuições	Modelos híbridos, workloads

5. Modelo Híbrido de Amostragem

O **método híbrido de amostragem** combina **dois ou mais métodos de geração de amostras aleatórias** (como transformada inversa, aceitação–rejeição e composição) para aproveitar as vantagens de cada um e contornar suas limitações.

Ideia principal

Nem todas as distribuições têm uma forma simples para ($F^{-1}(x)$) (inversa da CDF) ou uma função de densidade ($f(x)$) que facilite o uso de um único método.

O método híbrido busca **dividir o domínio ou estrutura da distribuição** e **usar o melhor método em cada parte**.

Como funciona

1. **Identificação das regiões ou componentes:**
   a. Partes da distribuição onde (F^{-1}) é simples → usa-se **Transformada Inversa**.
   b. Partes mais complexas → aplica-se **Aceitação–Rejeição** ou **Composição**.
2. **Combinação dos resultados:**
   a. As amostras geradas de cada parte são reunidas para formar um conjunto completo que segue a distribuição alvo.

Vantagens

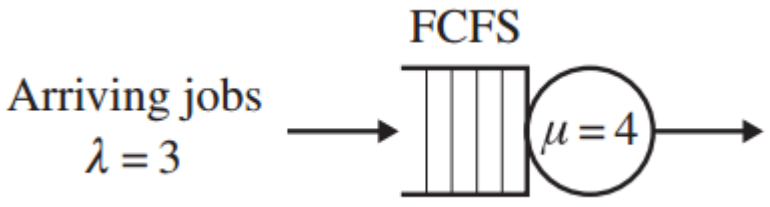
- Maior **eficiência** e **flexibilidade** que os métodos isolados.
- Permite tratar **distribuições complexas ou mistas** (contínuas e discretas, truncadas, ou multimodais).
- Reduz o número de rejeições e o custo computacional.

Exemplo típico

Para gerar amostras **Normais**, o método híbrido pode:

- Usar **Transformada Inversa** para a parte central da distribuição (onde (F^{-1}) é bem comportada);
- Usar **Aceitação–Rejeição** para as caudas (onde (F^{-1}) diverge).

6. Filas



Servidor: Qualquer recurso onde filas de tarefas possam se formar

Parâmetros do Sistemas

- Topologia da Rede

- Política (ordem de atendimento) da fila
- Average Arrival Rate:
    a. Taxa média de chegada por u.t.
- Mean interarrivel time: 1
- Size (s): Tamanho do job
    a. Tempo de serviço que o job demanda para ser concluído
- Mean Service Time: $E[S] = 1$
- Average Size Rate (): Taxa média nominal de serviço (em cada servidor) em jobs por u.t.

Métricas de Desempenho

- Response Time (**T**) : Tempo de resposta (Tempo no sistema)
  $T = Tq + Ts$
- Waiting Time (**Tq**) : Tempo perdido em filas (Q: Queue)
- Número de jobs no sistema (**N(t)**) no instante t
- Número de jobs na Fila (**Nq(t)**) no instante t

O tempo de serviço S, assim como outras V.A.s e métricas, **depende do servidor**. Será maior ou menor conforme a taxa de serviço do servidor onde está. Para referir-se às métricas do i-ésimo servidor em uma rede de filas, anota-se T_i , Tq_i , $N(t)_i$, etc.

Condição de Estabilidade: Sempre assumiremos que $<$

- **Tempo de Espera**
  $E[T] = E[Tq] + E[S]$
  Para $S \rightarrow$ Tempo de Processo
- **Número de Jobs no Sistema (N):** O número de jobs na fila mais as que estão sendo atendidas.
- **Número de Jobs na Fila (N_Q):** Apenas o número de jobs que estão esperando na fila.
- **Utilização (ρ_i):** A fração do tempo em que um dispositivo (servidor) `i` está ocupado. Em um sistema de servidor único, é calculada como $\rho_i = X_{ii}$
- **Vazão (Throughput: X_i):** A taxa de conclusão de jobs em um dispositivo `i` (jobs/segundo). Para um sistema estável, a taxa de saída é igual à taxa de entrada. $X_i = {}_i \cdot \rho_i$

Lei da Utilização: relaciona essas duas últimas métricas

$$\rho = \frac{X_i}{i}$$

Classificação das Redes de Filas

As redes de filas são geralmente classificadas em duas categorias principais:

1. **Redes Abertas (Open Networks):** Possuem chegadas e partidas externas.
   Vazão máxima é sempre limitada por
   Vazão real é igual a taxa de chegada (supondo $<$), ou seja, X_i não depende da taxa de serviço
2. **Redes Fechadas (Closed Networks):** Não possuem chegadas ou partidas externas.
   Um número fixo de jobs (N), conhecido como **nível de multiprogramação (MPL)**, circula constantemente pelo sistema. Elas se subdividem em:
   a. **Sistemas em Lote (Batch Systems):** Assim que uma tarefa termina, uma nova é iniciada imediatamente, mantendo sempre N jobs ativas no sistema.
   b. **Sistemas Interativos:** Modelam usuários em terminais. Um usuário envia uma requisição, espera pela resposta (tempo de resposta, R) e então passa um tempo "pensando" (think time, Z) antes de enviar a próxima requisição.

Lei de Little

Ela estabelece uma relação fundamental e simples entre o número médio de jobs em um sistema, a taxa de chegada e o tempo médio que uma tarefa passa no sistema.

$$E[X] = X \cdot E[T] \quad E[T] = \frac{1}{x} \cdot E[N]$$

Para $X =$ Vazão (Taxa média de jobs finalizados)
$\frac{1}{X} =$ Tempo entre jobs

Para sistemas abertos

$$E[N] = \ \cdot E[T]$$

Onde:

- **E[N]**: É o número médio de jobs no sistema (na fila + em serviço).
- **λ**: É a taxa média de chegada de jobs ao sistema.
- **E[T]**: É o tempo médio que um job passa no sistema (tempo de resposta ou *sojourn time*).

Para sistemas fechados

$$N = X \cdot E[T]$$

Onde:

- **N**: É o número de jobs no sistema, também conhecido como nível de multiprogramação (MPL).
- **X**: É a vazão (*throughput*) do sistema, ou seja, a taxa de conclusão de jobs.
- **E[T]**: É o tempo médio que uma tarefa leva para completar um ciclo no sistema. Para sistemas interativos, este tempo inclui o "tempo de pensamento" do usuário ($E[T] = E[R] + E[Z]$)

Apenas para a Fila

A lei também se aplica se considerarmos apenas a parte da fila do sistema:

$$E[N] = \ \cdot E[T]$$

Onde:

- N é o número de jobs na fila
- T é o tempo de espera na fila.

Para recortes do sistema

A Lei de Little pode ser usada para analisar apenas a parte da fila de um sistema, ignorando o tempo em que uma tarefa está sendo efetivamente servida.

A fórmula se torna: $E[N] = \ \cdot E[T]$

Onde:

- $E[N]$: O número médio de tarefas esperando na fila.
- : A taxa média de chegada de tarefas ao sistema.
- $E[T]$: O tempo médio que uma tarefa passa esperando na fila.

Lei dos Fluxos Forçados

A Lei dos Fluxos Forçados é uma lei operacional que estabelece uma relação direta entre a vazão (*throughput*) de um sistema inteiro e a vazão de um dispositivo individual dentro desse sistema.

$$X_i = E[V_i] \cdot X$$

Onde:

- X_i: É a vazão no dispositivo `i` (a taxa de conclusões de tarefas no dispositivo `i`).
- $E[V_i]$: É o número médio de visitas que uma tarefa faz ao dispositivo `i` antes de sair do sistema.
- **X**: É a vazão total do sistema (a taxa de conclusão de tarefas para o sistema como um todo)

Lei da Utilização:

A Lei de Little pode ser usada para provar a **Lei da Utilização**, que afirma que a utilização (ρi) de um servidor `i` é o produto de sua vazão (X_i) e o tempo médio de serviço ($E[S_i]$):

$$\rho_i = X_i \cdot E[S_i]$$

Lei do Gargalo

A Lei do Gargalo é uma lei operacional simples e poderosa usada para identificar o recurso que limita o desempenho de um sistema de filas. O "gargalo" do sistema é o dispositivo que possui a maior demanda total de serviço por tarefa.

Demanda de Serviço (Di)

Para entender a Lei do Gargalo, primeiro definimos a **demanda total de serviço (Di)** em um dispositivo `i`. Esta é a soma do tempo de serviço total que uma única tarefa exige do dispositivo `i` em todas as suas visitas

$$E[D_i] = E[V_i] \cdot E[S_i]$$

Onde:

- $E[V_i]$: O número médio de visitas que uma tarefa faz ao dispositivo `i`.
- $E[S_i]$: O tempo médio de serviço no dispositivo `i` por visita.

Para um sistema real

$$E[D_i] = \frac{B_i}{C}$$

Onde,

- B_i : Tempo total que o servidor ficou ocupado
- C : Total de jobs finalizados pelo sistema

A Lei do Gargalo

A lei estabelece uma relação direta entre a utilização de um dispositivo, a vazão do sistema e a demanda de serviço:

$$\rho_i = X \cdot E[D_i]$$

Onde:

- **ρi**: É a utilização do dispositivo `i` (a fração de tempo que ele está ocupado).
- **X**: É a vazão (*throughput*) total do sistema (jobs concluídos por segundo).
- $E[D_i]$: É a demanda média de serviço total no dispositivo `i`.

Identificando o gargalo

O dispositivo com a maior demanda de serviço total,

$$D_{max} = \max_i E[D_i]$$

é o **dispositivo gargalo**. Este dispositivo é o principal fator que limita o desempenho geral do sistema, pois é o primeiro a atingir 100% de utilização à medida que a carga aumenta.

7. Análises de Modificações de Sistemas Fechados

Métricas principais

Em sistemas fechados, analisam-se:

- **Throughput (X)** – taxa de processamento de requisições no sistema;
- **Tempo médio de resposta (E[R])** – tempo total no sistema (espera + serviço);
- **Número médio de clientes (N)** – total constante;
- **Tempo de reflexão (E[Z])** – tempo médio que o cliente passa (ocsioso) fora do sistema, antes de voltar.

Conceitos-Chave

A análise se baseia em **limites assintóticos** para a vazão (*throughput*, X) e o tempo de resposta (E[R]) em sistemas fechados. Esses limites são definidos em termos da demanda total de serviço em cada dispositivo (D_i).

- **Dispositivo Gargalo (D_{max})** É o dispositivo com a maior demanda total de serviço por tarefa. Este é o recurso que fundamentalmente limita o desempenho do sistema.

$$D_{max} = \max_i E[D_i]$$

- **Soma das Demandas (D):** A soma das demandas média de serviço em todos os dispositivos (ou seja, demanda total no sistema)

$$D = \sum_i E[D_i]$$

  Os limites para a vazão (X) e o tempo de resposta (E[R]) em um sistema com N jobs são dados por:

$$X \leq \min\left(\frac{N}{D + E[Z]}, \frac{1}{D_{max}}\right) \quad (1)$$

  .

$$E[T] = E[R] + E[Z] \geq D + E[Z] \quad (2)$$

Onde

- **E[T]**: É o **tempo médio de ciclo do sistema**.
- **E[R]**: É o **tempo médio de resposta**. É o tempo que o sistema leva para processar uma tarefa
- **E[Z]**: É o **tempo médio de reflexão**, que é zero para sistemas em lote
- N: Número total de cliente

- D: É a a soma das demandas de serviço médias em todos os dispositivos para uma única tarefa
- D_{max}: É o dispositivo com a maior demanda

(2) significa que o tempo médio de resposta que os usuários estão experimentando é **maior do que o tempo mínimo de serviço**
Pela lei de little:

$$N = X \cdot E[T] \longrightarrow N = X \cdot E[R] + E[Z]$$

Por $E[R] \geq D$:

$$X \leq \frac{N}{E[R] + E[Z]} \longrightarrow X \leq \frac{N}{D + E[Z]}$$

Pela de lei do gargalo (a utilização ρ_i deve ser menor que 100%)

$$X \cdot E[D_i] \leq 1 \longrightarrow X \leq \frac{1}{E[D_i]}$$

Como isso chegamos em (1), o **limite superior assintótico** para a vazão (X)

- **Ponto de Saturação (N)** É o nível de multiprogramação (número de jobs) além do qual começa a haver enfileiramento significativo no sistema.

$$\frac{N}{D + E[Z]} = \frac{1}{D_{max}}$$

$$N = \frac{D + E[Z]}{D_{max}}$$

Principais Conclusões da Análise

A análise desses limites revela como o sistema se comporta sob diferentes níveis de carga:

1. **Quando a Carga é Alta (N N):** O desempenho (vazão e tempo de resposta) é dominado pelo dispositivo gargalo (D_{max}) $\frac{1}{D_{max}}$
   Para obter mais vazão ou menor tempo de resposta, é necessário diminuir a demanda no dispositivo gargalo. Melhorar qualquer outro dispositivo terá um efeito insignificante no desempenho geral.
2. **Quando a Carga é Baixa ($N < N$):** O desempenho é limitado pela soma total das demandas (D). $\frac{N}{D+EZ}$
   Melhorar qualquer dispositivo (não apenas o gargalo) pode levar a uma pequena melhoria no desempenho.

Comparação entre Sistemas Fechados e Abertos

A análise de modificações se aplica de forma diferente a sistemas abertos e fechados:

- **Sistemas Fechados:** Como visto, o desempenho em alta carga é rigidamente limitado pelo gargalo.
- **Sistemas Abertos:** A vazão é determinada pela taxa de chegada externa ( ). Embora a vazão ainda seja limitada pela capacidade do gargalo ($X \leq 1 D_{max}$), essa não é uma restrição tão forte quanto nos sistemas fechados. Em um sistema aberto, melhorar um dispositivo que não é o gargalo **ainda assim melhora o tempo de resposta médio**, ao contrário do que acontece em um sistema fechado sob alta carga.