Capítulo 2

Distribuições de Probabilidade

2.1 Introdução

UMa das abordagens mais úteis para modelagem de sistemas é a caracterização de eventos através de distribuições de probabilidade. As distribuições contem muitas informações acerca de um certo evento e podem ser utilizadas para realizar a análise de desempenho utilizando abordagens analíticas ou através de simulações. O objetivo deste capítulo é introduzir ao leitor os principais conceitos necessários para utilizar distribuições de probabilidade na modelagem de sistemas.

2.2 Eventos e Probabilidade

Sejam A e B eventos. O evento de pelo menos um entre A ou B ocorra pode ser escrito por $A \cup B$. O evento de ambos ocorrerem é chamado interseção de A e B, escrito por $A \cap B$ (ou simplesmente AB).

A probabilidade de um evento A é escrita por $P(A)$. Um evento C que *sempre* ocorre possui $P(C) = 1$. O evento impossível $\varnothing$ tem $P(\varnothing) = 0$. Eventos A e B são disjuntos se $A \cap B = \varnothing$. Para eventos disjuntos, (ou independentes), pode-se escrever $P(A \cup B) = P(A) + P(B)$. Como caso geral, pode-se escrever

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2.1}$$

Seja ε um experimento e S o espaço amostral associado a ele. A cada evento s, $s \in S$, associaremos um número real representado por $P(s)$, denominado probabilidade de ocorrência de s, que satisfaça as seguintes propriedades: $0 \leq P(s) \leq 1$ e $P(S) = 1$;

2.3 Variáveis Aleatórias

Considere novamente ε um experimento e S o seu espaço amostral. Uma função X, que associe a cada elemento de $s \in S$ um número real, $X(s)$, ou simplesmente X, é denominada variável aleatória. Um exemplo consiste em supor como variável aleatória X o número de *coroas* obtidos no lançamento de duas moedas - os resultados possíveis são {cara, cara},{cara, coroa}, {coroa, cara} e {coroa, coroa}, e o espaço amostral possível será $S = \{0, 1, 2\}$.

Normalmente, são utilizadas letras maiúsculas para denotar variáveis aleatórias e letras minúsculas para números reais. Por exemplo, suponha que foi observado o tráfego em um servidor de páginas (este é o experimento). Pode-se identificar as seguintes variáveis aleatórias que descrevem alguns aspectos do experimento:

Tamanho do pacote (Y) : espaço amostral consiste dos números inteiros maiores ou iguais a zero ($0 \leq y < \infty$, $y \in \mathbb{Z}$);

Intervalo de tempo entre pacotes (W) : espaço amostral consiste dos números reais maiores que zero ($0 < w < \infty$, $w \in \mathbb{R}$);

Tempo de atendimento da requisição (Z) : espaço amostral consiste dos números reais maiores que zero ($0 < z < \infty$, $z \in \mathbb{R}$);.

A expressão $\{X = x\}$ indica que a variável randômica assume um valor igual ao número x. A probabilidade de ocorrência de x é denotada por $P(X = x)$. Permitindo a variação de x, obtém-se a *função densidade de probabilidade* da variável aleatória X, normalmente denotada por

$$f(x) = P(X = x), -\infty < x < +\infty \tag{2.2}$$

A expressão $(X \leq x)$ indica que a variável aleatória assume um valor menor ou igual a x. Permitindo a variação de x, obtém-se a função de distribuição acumulada de probabilidade da variável X, que pode ser escrita por

$$F(x) = P(X \leq x), -\infty < x < +\infty \tag{2.3}$$

A *função densidade de probabilidade* e está relacionada com a *distribuição acumulada de probabilidade* da seguinte maneira:

$$F(x) = \int_{-\infty}^{x} f(x)dx, -\infty < x < +\infty \tag{2.4}$$

$$f(x) = \frac{d}{dx}F(x), -\infty < x < +\infty \tag{2.5}$$

Suponha que X seja uma variável aleatória contínua e existe uma função de densidade de probabilidade $f(x)$, definida para todos os valores de x, $a \leq x \leq b$, com $a \leq b$. Neste caso, pode-se escrever as seguintes propriedades:

$$\begin{cases} f(x) \geq 0, & \text{para todo x} \\ \int_a^b f(x)dx = 1 \end{cases} \tag{2.6}$$

Uma distribuição de probabilidade é chamada *discreta* caso exista um conjunto enumerável de valores $x_1, x_2, \ldots, x_n$ tais que $P(X = x_i) > 0$, de forma que $\sum_{i=1}^{n} f(x_i) = 1$, sendo n o número de valores discretos possíveis. Par uma variável discreta, as funções densidade de probabilidade e probabilidade acumulada estão relacionadas por

$$F(x_i) = \sum_{j=1}^{j=i} f(x_j) \tag{2.7}$$

2.3.1 Momentos, Esperança e Variância

O enésimo momento de uma variável aleatória X é dado por

$$E[X^n] = \int_{-\infty}^{+\infty} x^n f(x)dx \tag{2.8}$$

Esta equação descreve o *momento central em torno de zero*, e pode ser também escrito como $E[(X-0)^n]$. O $n-$ésimo momento central em torno de E[X] é dado por

$$E[X - E[X]]^n \tag{2.9}$$

O primeiro momento central em torno de zero, correspondendo a $n = 1$ é chamado de valor esperado (ou média) de X, e normalmente denotado por $E[X]$ (e muitas vezes por μ_x). O primeiro momento central em torno da média é zero, pois fazendo $n = 1$ na equação 2.9 obtemos

$$\begin{aligned} E[X] &= E[X - E[X]] \\ &= E[X] - E[E[X]] \\ &= E[X] - E[X] \\ &= 0 \end{aligned}$$

O valor esperado de uma variável aleatória X significa que, se tomarmos uma quantidade muito grande de determinações de X, a média destes valores converge para $E[X]$. O leitor deve

ter cuidado no uso de $E[X]$ para o dimensionamento de sistemas; por exemplo, suponha que a variável aleatória X representa do arquivo transmitido por um servidor Web. A variação do valor de X em torno da média é uma informação importante e não é representado por $E[X]$.

O segundo momento central *em torno da média* é chamado de *variância* de X, denotado normalmente como $V[X]$ ou σ^2 e dado por

$$\begin{aligned} V[X] &= E[(X - E[X])^2] \\ &= \int_{-\infty}^{+\infty} (x - E[X])^2 f(x) dx \end{aligned}$$

Normalmente σ denota o desvio padrão de X, dado por $\sigma = \sqrt{V[X]}$. A variância fornece uma estimativa sobre a variação dos valores observados em relação à esperança. No exemplo considerado anteriormente, onde X representa o tamanho do arquivo transmitido por um servidor, caso todos os valores observados sejam exatamente iguais à média, então $V[X] = 0$. Conforme a aumenta a diferença entre os valores observados e a esperança, o valor de $V[X]$ também aumenta. Pode-se escrever a variância em função da esperança, da seguinte forma:

$$\begin{aligned} V[X] &= E[(X - E[X])^2] \\ &= E[(X^2 - 2XE[X] + E[X]^2)] \\ &= E[X^2] - E[2XE[X]] + E[X]^2 \\ &= E[X^2] - 2E[X]E[E[X]] + E[X]^2 \\ &= E[X^2] - 2E[X]^2 + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Exemplo 1: Suponha uma variável aleatória X com função densidade de probabilidade dada por $f(x) = 3x^2/a^3$, definida para $0 \leq x < a$. A esperança $E[X]$ pode ser calculada através de $E[X] = \int_0^a x(3x^2/a^3)dx$, que tem como resultado $E[X] = 3a/4$. A variância de $V[X]$ é dada por $V[X] = \int_0^a (x - E[X])^2 3x^2/a^3 dx$, que nos leva a $V[X] = \frac{3}{80}a^2$.

□

2.3.2 Coeficiente de Correlação

Suponha duas variáveis aleatórias unidimensionais, denotadas por X e Y. O coeficiente de correlação mede o grau de associação entre X e Y, dado por

$$\rho_{xy} = \frac{E[(X - E[X])(Y - E[Y]]}{\sqrt{V[X]V[Y}}, \text{ ou } \rho_{xy} = \frac{E[XY] - E[X]E[Y]}{\sqrt{V[X]V[Y]}} \tag{2.10}$$

O valor de ρ_{xy} está no intervalo entre -1 e 1. Se X e Y forem independentes, então $\rho_{xy} = 0$, de forma que quanto mais próximo $|\rho_{xy}|$ de um maior o grau de associação entre X e Y.

O numerador de ρ_{xy} é chamado de covariância entre X e Y, denotado por $COV[X, Y] = E[(X - E[X])(Y - E[Y]]$.

2.4 Principais distribuições de probabilidade

2.4.1 Principais Distribuições Contínuas

Distribuição Uniforme

A distribuição uniforme é uma das mais simples disponíveis. A probabilidade de ocorrência de um evento é constante no intervalo a, b. A função densidade de probabilidade e detalhes da distribuição uniforme são apresentados na Tabela 2.1.

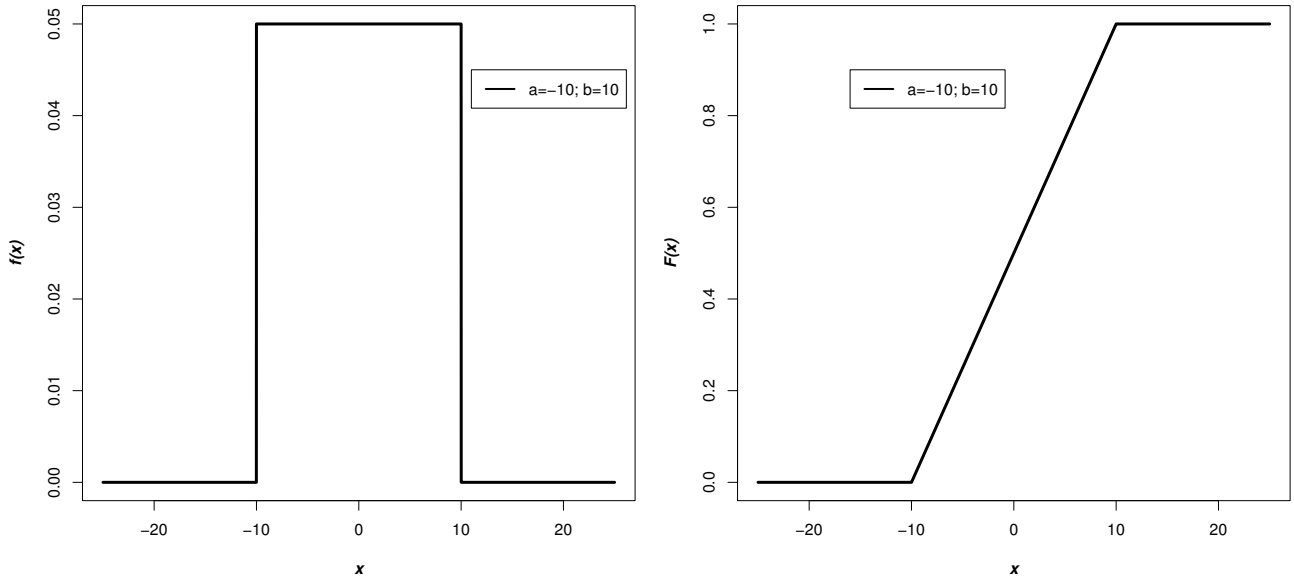


Figura 2.1: Distribuição uniforme

Tabela 2.1: Principais distribuições de probabilidade

Distribuição Uniforme	
Parâmetros	a, b $a =$ limite inferior $b =$ limite superior
Limites	$a \leq x \leq b$
Densidade de Probabilidade	$f(x) = \frac{1}{b-a}$
Distribuição Acumulada	$F(x) = \frac{x-a}{b-a}$ se $a \leq x < b$ 0 se $x < a$ 1 se $x \leq b$
Esperança ($E[X]$)	$\mu = \frac{a+b}{2}$
Variança ($Var[X]$)	$\sigma^2 = \frac{(b-a)^2}{12}$

Distribuição Exponencial	
Parâmetros	μ $\mu =$ média, $\mu > 0$
Limites	$0 \leq x < \infty$
Densidade de Probabilidade	$f(x) = \frac{1}{\mu} e^{-x/\mu}$
Distribuição Acumulada	$F(x) = 1 - e^{-x/\mu}$
Esperança ($E[X]$)	μ
Variança ($Var[X]$)	μ^2

Distribuição Normal	
Parâmetros	μ, σ $\mu =$ média $\sigma =$ desvio padrão, $\sigma > 0$
Limites	$-\infty < x < +\infty$
Densidade de Probabilidade	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
Esperança ($E[X]$)	μ
Variança ($Var[X]$)	σ^2

Distribuição Lognormal	
Parâmetros	μ, σ $\mu =$ média de ln(x), $\mu > 0$ $\sigma =$ desvio padrão de ln(x), $\sigma > 0$
Limites	$0 < x < +\infty$
Densidade de Probabilidade	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(lnx-\mu)^2/2\sigma^2}$
Esperança ($E[X]$)	$e^{\mu+\sigma^2/2}$
Variança ($Var[X]$)	$e^{\mu+\sigma^2/2}(e^{\sigma^2} - 1)$

Distribuição de Weibull	
Parâmetros	α, β $\alpha > 0, \alpha =$ parâmetro de forma $\beta > 0, \beta =$ parâmetro de escala
Limites	$0 \leq x < +\infty$
Densidade de Probabilidade	$f(x) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha} e^{-(x/\beta)^\alpha}$
Distribuição Acumulada	$F(x) = 1 - e^{-(x/\beta)^\alpha}$
Esperança ($E[X]$)	$\frac{\beta}{\alpha}\Gamma(1/\alpha)$ $\Gamma(b) = \int_0^{+\infty} e^{-x} x^{b-1} f(x) dx$
Variança ($Var[X]$)	$\frac{\beta^2}{\alpha^2}(2\alpha\Gamma(2/\alpha) - [\Gamma(1/\alpha)]^2$

Distribuição de Pareto	
Parâmetros	α, β $\alpha > 0$, parâmetro de forma $\beta > 0$, parâmetro de escala
Limites	$b \leq x < +\infty$
Densidade de Probabilidade	$f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}$
Distribuição Acumulada	$F(x) = 1 - (\frac{\beta}{x})^\alpha$
Esperança ($E[X]$)	$\frac{\alpha\beta}{\alpha-1}$, $\alpha > 1$
Variança ($Var[X]$)	$\frac{\alpha\beta}{(\alpha-1)^2(\alpha-2)}$, $\alpha > 2$

Distribuição Binomial	
Parâmetros	p, n p = prob. sucesso em uma tentativa $n =$ número de repetições
Limites	$x = 0, 1, 2, \ldots, \infty$
Densidade de Probabilidade	$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$
Esperança ($E[X]$)	np
Variança ($Var[X]$)	$np(1-p)$

Distribuição de Poisson	
Parâmetros	λ $\lambda =$ média
Limites	$x = 0, 1, 2, \ldots, \infty$
Densidade de Probabilidade	$f(x) = \lambda^x \frac{e^{-\lambda}}{x!}$
Esperança ($E[X]$)	λ
Variança ($Var[X]$)	λ

Distribuição de Geométrica	
Parâmetros	p $p =$ prob. de sucesso, $0 < p < 1$
Limites	$x = 0, 1, 2, \ldots, \infty$
Densidade de Probabilidade	$f(x) = (1-p)^{(x-1)} p$
Esperança ($E[X]$)	$1/p$
Variança ($Var[X]$)	$\frac{1-p}{p^2}$

Distribuição Triangular	
Parâmetros	a, b, m
Limites	$a \leq x \leq b$, $a < m$, $m < b$
Densidade de Probabilidade	$f(x) = \frac{2(x-a)}{(m-a)(b-a)}$ para $a \leq x < m$ $f(x) = \frac{2(b-x)}{(b-m)(b-a)}$ para $m \leq x \leq b$
Esperança ($E[X]$)	$\frac{a+b+m}{3}$
Variança ($Var[X]$)	$\frac{a(a-m)+b(b-a)+m(m-b)}{18}$

Distribuição Exponencial

A distribuição exponencial pode modelar muitos eventos que ocorrem na natureza. A distribuição exponencial não possui memória: a ocorrência de um evento não está relacionado com o eventos passados. Normalmente, o intervalo de tempo entre: chegadas de chamadas VoIP, requisições Web ou mesmo de usuários em filas de banco são modeladas por uma distribuição exponencial.

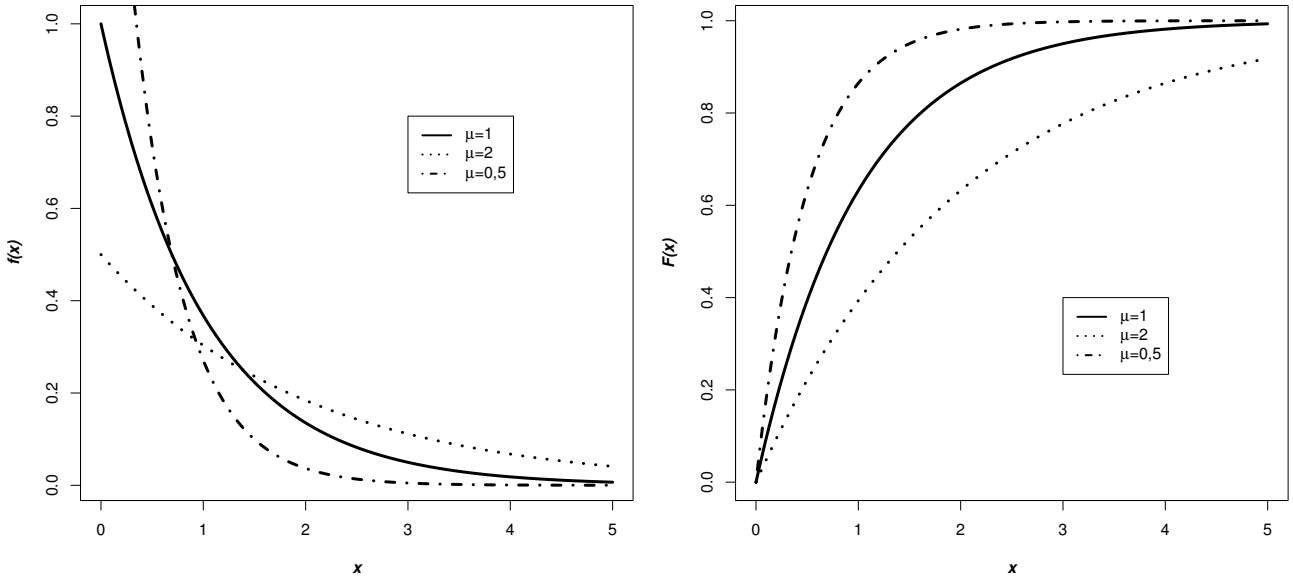


Figura 2.2: Distribuição Exponencial

Distribuição Normal

A distribuição normal (ou *gaussiana*) é a mais proeminente distribuição em estatística e probabilidade. O seu grande uso justifica-se pelo teorema do limite central, que estabelece que a distribuição da média ou da soma de i observações independentes, de qualquer distribuição, se aproxima da distribuição normal quando $i \to \infty$. Note que a distribuição normal não possui expressão analítica conhecida para $F(x)$. A falta de expressão analítica para $F(x)$ irá causar um problema com a geração de variáveis aleatórias com distribução normal para simulações, uma vez que um dos métodos mais utilizados precisa da expressão para $F(x)$, conforme será tratado posteriormente.

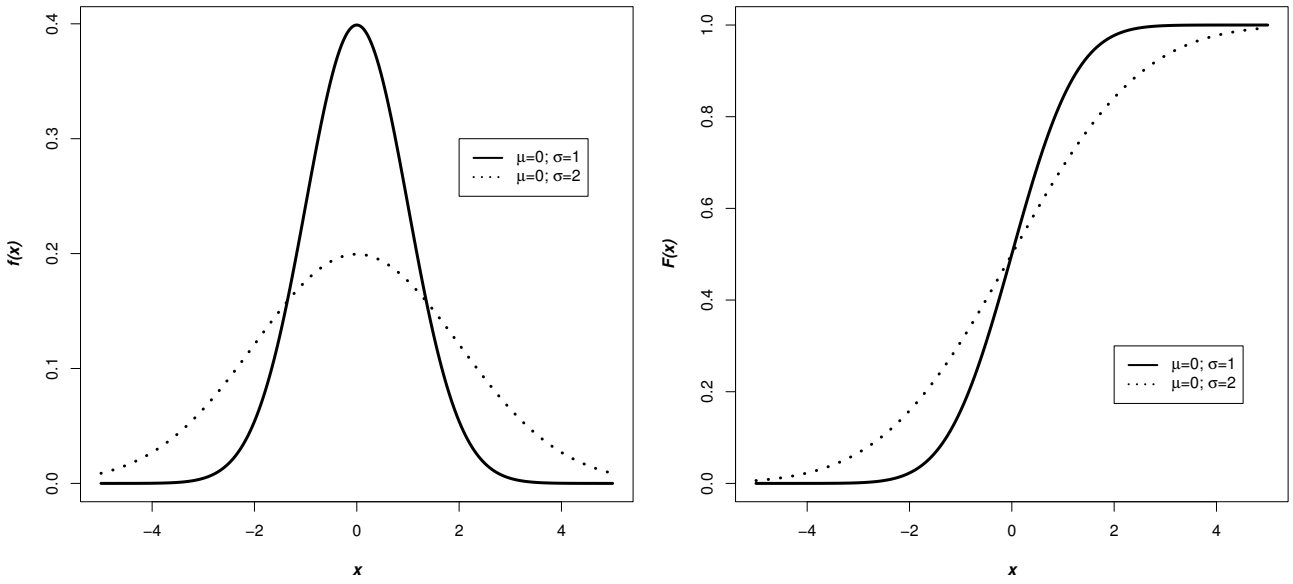


Figura 2.3: Distribuição Normal

Distribuição Lognormal

A distribuição lognormal é uma distribuição de decaimento semi-exponencial, o que significa que a convergência da densidade de probabilidade é mais lenta que a distribuição exponencial mas não é tão lenta quanto as distribuições de cauda pesada. Caso os dados sejam transformados utilizando uma função logaritmica, $y = log(x)$, caso y apresente distribuição normal, x terá distribuição lognormal.

Como exemplo de aplicação, tem sido reportado que o tamanho dos objetos transmitidos em por servidores Web pode ser modelados habitualmente por uma distribuição lognormal.

Distribuição de Weibull

A distribuição de Weibull também é uma distribuição de decaimento semi-exponencial. A distribuição exponencial é um caso específico da distribuição de Weibull para $\beta = 1$.

Tipicamente, a distribuição de Weibull é utilizada na modelagem do tempo de vida de componentes; para $\alpha < 1$ a distribuição fornece uma taxa de falhas que aumenta ao longo do tempo; com $\alpha = 1$, a taxa de falhas é constante ao longo do tempo e com $\alpha > 1$ a taxa de falhas reduz ao longo do tempo.

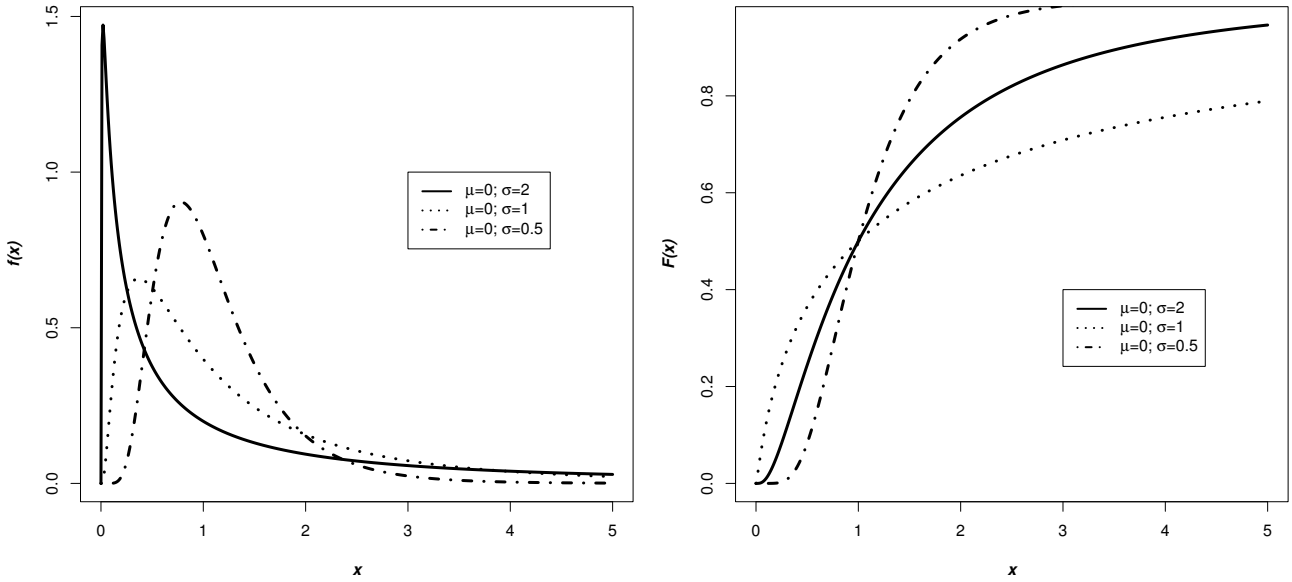


Figura 2.4: Distribuição lognormal

Distribuição de Pareto

A distribuição de Pareto é uma distribuição de cauda pesada cuja função de distribuição de probabilidade e demais características apresentadas na Tabela 2.1, onde α é o parâmetro de forma e β é chamado de parâmetro de escala. A principal característica de uma variável aleatória como esta é a variabilidade extrema exibida. Note que a média converge apenas para valores de $\alpha > 1$ e a variância para $\alpha > 2$. Para $\alpha > 2$, quanto mais próximo de 2 estiver $alpha$, maior será variância exibida.

Distribuições de cauda pesada tem sido utilizadas para modelar diversos fenômenos em telecomunicações, como a duração de chamadas telefônicas VoIP (Voz sobre IP), onde o parâmetro α típico está entre 2,2 e 2,8.

2.4.2 Principais Distribuições Discretas

Distribuição Binomial

O número de sucessos x em uma seqüência de n repetições possui uma distribuição binomial, sendo p a probabilidade de sucesso em uma única tentativa. As principais características da distribuição binomial são apresentadas na Tabela 2.1.

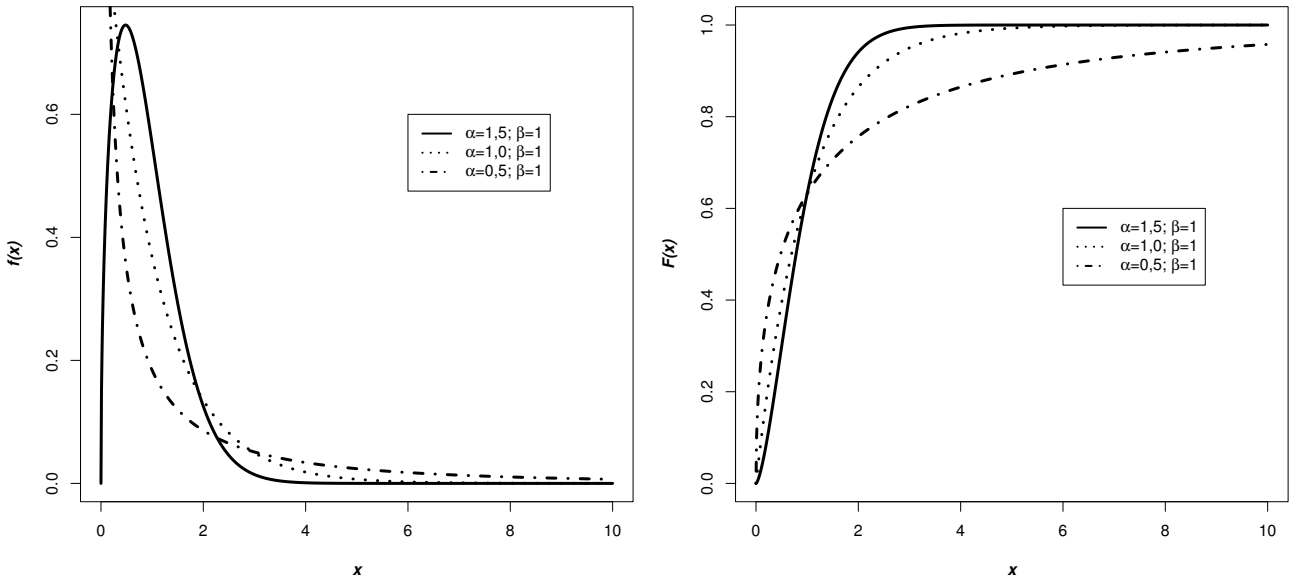


Figura 2.5: Distribuição de Weibull

Distribuição Geométrica

A distribuição geométrica é o equivalente discreto à distribuição exponencial, e como esta, não possui memória. Os usos desta distribuição são tipicamente ...

Distribuição de Poisson

A distribução de Poisson é utilizada intensamente em modelos de filas, uma vez que é comum que se possa caracterizar o número de chegadas em um determinado intervalo de tempo através desta distribuição. Uma propriedade particularmente interessate é a que permite que um grade número de fontes independentes que geram eventos com a distribuição de Poisson pode ser estudada como sendo uma fonte única onde o parâmetro λ resultante é a soma das médias de cada uma das fontes.

Também é digno de nota que para médias grandes torna-se difícil calcular o valor da distribuição; neste caso, uma boa aproximação pode ser obtida através do uso da distribuição normal.

2.5 Distribuições de Cauda Pesada

Uma variável aleatória X possui distribuição de cauda pesada se:

$$P(X > x) \sim c.x^{-\alpha}, \quad x \to \infty \tag{2.11}$$

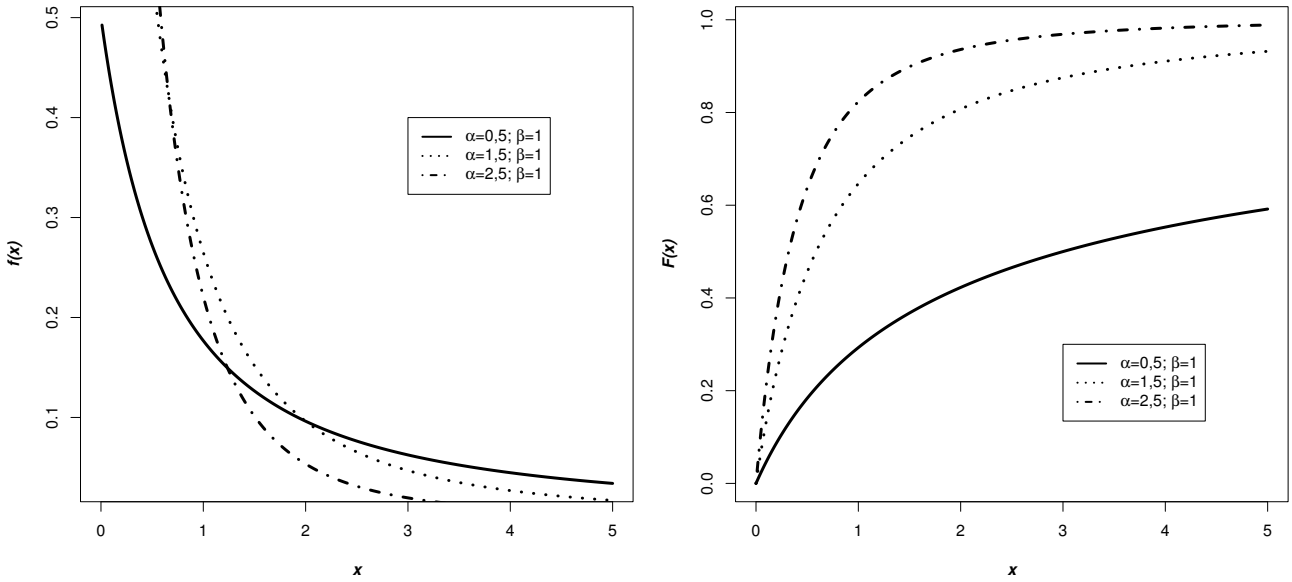


Figura 2.6: Distribuição de Pareto

onde α é chamado parâmetro de forma e c é uma constante positiva. Distribuições de cauda pesada possuem uma série de propriedades que as diferenciam das distribuições mais comuns (como a distribuição exponencial ou normal). A principal característica de uma variável aleatória com estas características é a variabilidade extrema exibida.

A literatura sugere que a auto-similaridade do tráfego observado em redes de telecomunicações está associada ao tamanho dos objetos transmitidos, que possuem frequentemente características de cauda pesada. Uma das maneiras utilizadas para verificar se uma distribuição possui cauda pesada é através do gráfico da distribuição complementar em escala logarítmica (*log-log complementary distribution, LLCD*). Este gráfico representa o valor da distribuição de probabilidade acumulada complementar $\overline{F}(x) = 1 - F(x)$ com $F(x) = \Pr\{X <= x\}$, em escala logarítmica, de modo a procurar a invariância dada por:

$$\frac{d\,log(\overline{F}(x))}{d\,log(x)} = -\alpha\, x > \theta \tag{2.12}$$

Na prática, escolhe-se um valor de θ a partir do qual o gráfico passa a ser linear em pelo menos 2 décadas.

Exemplo 2.1 *Como exemplo, pode ser estudado o tempo de duração de uma chamada telefônica no sistema VoIP (Voz sobre IP). A Figura 2.8 mostra que o gráfico da distribuição complementar comporta-se como uma reta na escala logarítmica a partir de do tempo de duração da sessão de aproximadamente* 100 *segundos, o que indica a presença da cauda pesada. Grafica-*

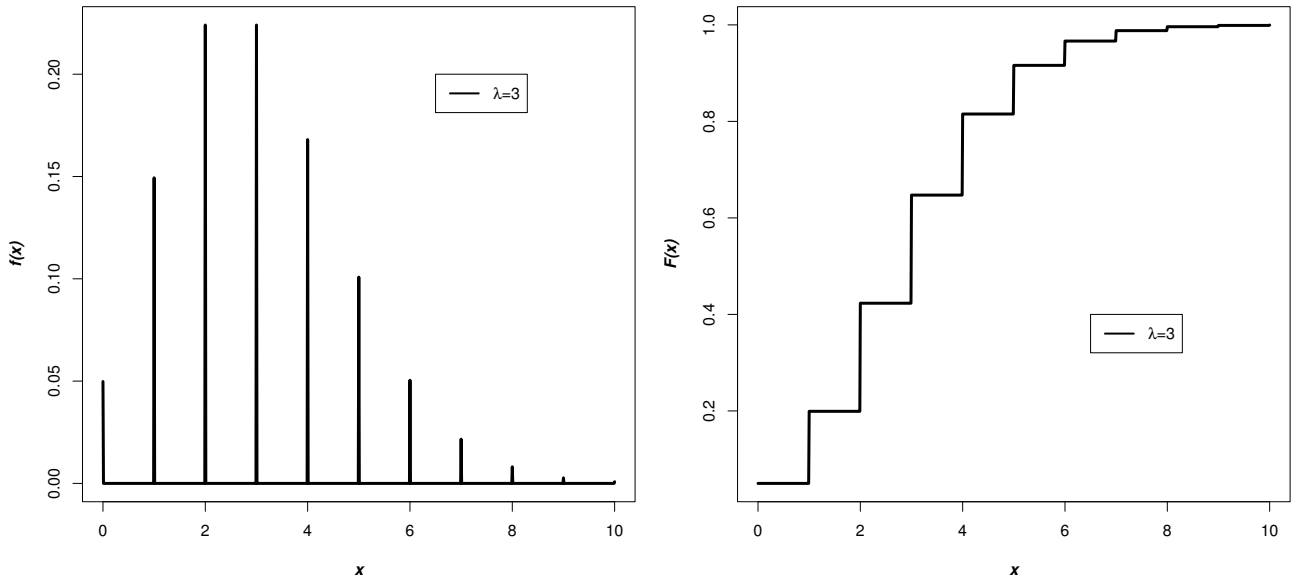


Figura 2.7: Distribuição de Poisson

mente pode ser constatado um ângulo de aproximadamente -63^o*, o que pode ser utilizado de acordo com a Equação 2.12 para estimar o valor do parâmetro de forma* $\widehat{\alpha} \simeq 1,96$*.*

2.6 Ajuste de distribuições

É muito interessante que uma certa variável aleatória em estudo possa ser modelada por uma determinada distribuição de probabilidade teórica. Isto é particularmente importante porque existem muitos resultados teóricos conhecidos utilizando as distribuições teóricas mais comuns. Além disso, as distribuições teóricas mais usadas apresentam em geral certas facilidades para o tratamento analítico. Desta forma, para que o analista possa utilizar este ferramental já desenvolvido, primeiro ele deve determinar se existe uma distribuição conhecida que pode ser utilizada para modelar os dados empíricos. Um exemplo de aplicação é o uso da teoria de filas: antes de iniciar a análise, devem ser caracterizados os processos de chegada e atendimento.

2.6.1 Histograma

O histograma é uma técnica de representação gráfica nos dá uma idéia sobre o formato da distribuição de probablidade dos dados em estudo, mostrando as frequências relativas no eixo das abcissas y e classes (ou agrupamentos) no eixo das ordenadas x.

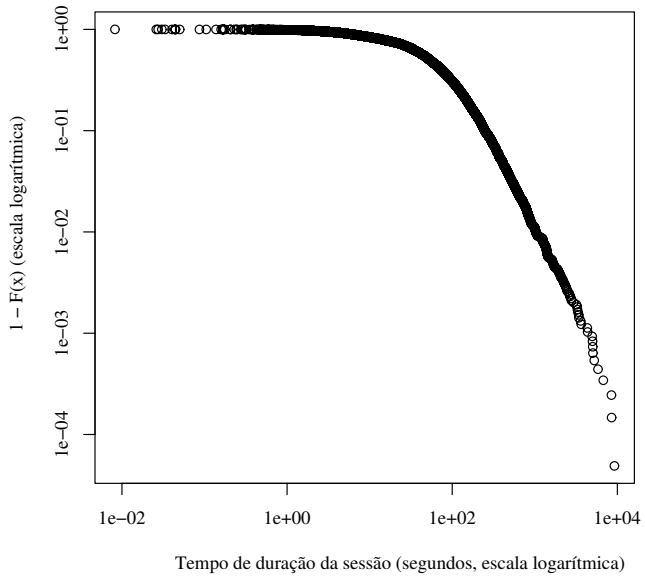


Figura 2.8: Distribuição complementar em escala logarítmica para o tempo de duração da sessão VoIP amostrado em uma grande operadora de telecomunicações no Brasil

O primeiro passo para a criação do histograma é montar uma tabela de frequência, identificando o número de classes e seus limites. Uma classe agrupa as observações dentro de um certo limite de valores. Para obter um bom resultado, o número de classes a ser utilizado depende do número de pontos na amostra e da própria distribuição dos pontos ao longo dos limites inferior e superior dos dados. Um bom ponto de partida é iniciar com a raiz quadrada do número de elementos amostrados - no entanto, o analista pode obter resultados melhores observando o número de classes e o número de observações em cada classe - um grande número de observações em uma classe e um pequeno em outras indica que a divisão de classes deve ser refeita.

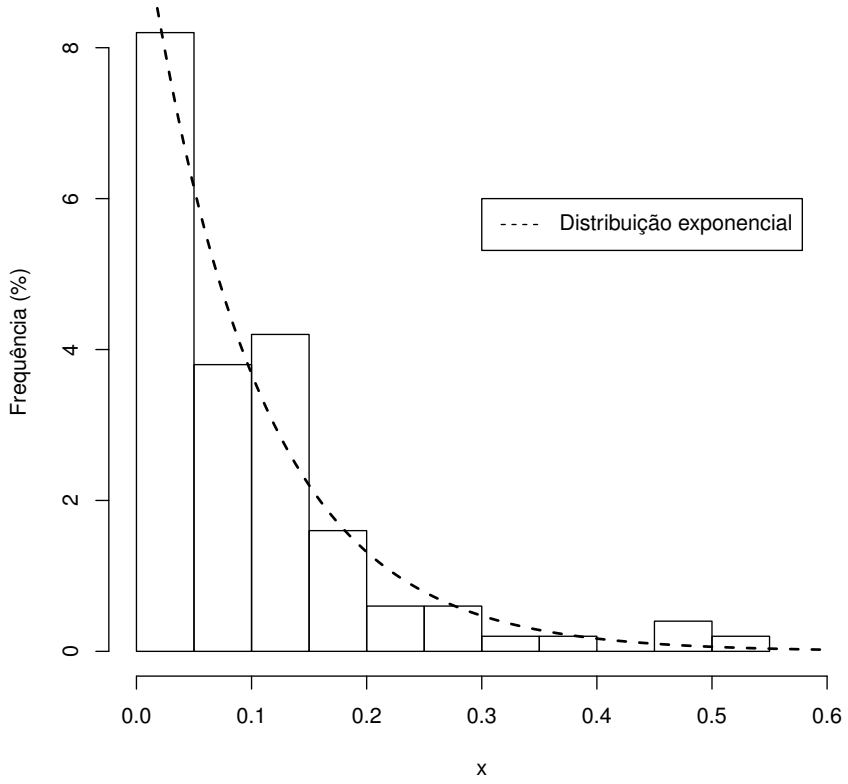
Exemplo 2.2 *Suponha que o intervalo de tempo entre chegada de requisições à um servidor*

Tabela 2.2: Intervalo de tempo entre chegadas de requisições em um servidor Web

0.5314	0.0357	0.0273	0.2188	0.2188	0.0788	0.1557	0.1605	0.0337	0.2607
0.2009	0.0366	0.2935	0.0237	0.4973	0.0306	0.0059	0.0403	0.0024	0.0187
0.1448	0.0479	0.0943	0.3216	0.0102	0.0109	0.0089	0.1096	0.0579	0.0102
0.1031	0.0613	0.0358	0.0579	0.4676	0.1101	0.0259	0.0112	0.0532	0.0018
0.1628	0.1193	0.0595	0.0141	0.1058	0.0277	0.1169	0.1221	0.1622	0.0345
0.0020	0.0823	0.0105	0.0622	0.0454	0.1983	0.0673	0.1023	0.1090	0.1952
0.0066	0.0281	0.0308	0.0286	0.1379	0.0624	0.1028	0.1088	0.1308	0.1700
0.0956	0.1246	0.0285	0.0166	0.1226	0.0325	0.0434	0.0699	0.2957	0.0754
0.0719	0.0473	0.1027	0.0327	0.0074	0.0603	0.0092	0.1291	0.0518	0.3773
0.0138	0.0095	0.0584	0.1092	0.0396	0.0716	0.0271	0.1024	0.1219	0.1712

Web é apresentado na Tabela 2.2. Para os dados apresentados, o valor médio pode ser estimado em $0,1014$ *e o desvio padrão em* $0,1260$*. O histograma resultante é apresentado a seguir:*

Intervalo	*Frequência observada*	*Intervalo*	*Frequência observada*
$0.00 \leq x < 0.05$	41	$0.25 \leq x < 0.30$	3
$0.05 \leq x < 0.10$	19	$0.30 \leq x < 0.35$	1
$0.10 \leq x < 0.15$	21	$0.35 \leq x < 0.40$	1
$0.15 \leq x < 0.20$	8	$0.40 \leq x < 0.45$	0
$0.20 \leq x < 0.25$	3	$0.45 \leq x < 0.50$	2



2.6.2 Testes de Aderência

O objetivo de um teste de aderência é verificar se, com um certo grau de confiança, uma distribuição teórica conhecida pode ser utilizada para modelar um conjunto de dados em estudo. Para realizar o teste de aderência, o analista deve formular uma hipótese sobre qual distribuição teórica será utlizada e seus parâmetros. O histograma pode fornecer uma idéia sobre a forma da distribuição, que pode ser comparada graficamente com uma distribuição conhecida. Os parâmetros da distribuição teórica são estimados de acordo com cada caso.

2.6.3 Teste de Aderência de Chi-Quadrado

O histograma é um estimador da função de densidade de probabilidade da amostra. Neste ponto, é necessário determinar se existe alguma distribuição clássica de probabilidade que pode ser ajustada à distribuição amostral. Para tal tarefa, a partir do histograma pode-se realizar um teste de aderência.

No caso do exemplo mostrado, o formato da distribuição é semelhante a *Distribuição Exponencial*.

Para realizar o teste de aderência de Chi-Quadrado, deve-se calcular:

$$D = \frac{(fo - fe)^2}{fe} \tag{2.13}$$

onde

$\mathbf{f}_o$ Freqüência observada na amostra;

$\mathbf{f}_e$ Freqüência esperada obtida a partir da distribuição teórica de probabilidade.

f_e pode ser obtido a partir da $\int_i^j f(x)dx$, com i e j identificando o limite inferior e superior da classe. $f(x)$ representa a distribuição de probabilidade alvo. Sempre que o número de ocorrências for pequeno, a classe deve ser agrupada com classes adjacentes - é recomendável, para aplicação deste teste, que a amostra possua pelo menos 25 elementos e que a freqüência esperada em cada classe seja maior que cinco.

Exemplo 2.3 *Utilizando os dados do Exemplo 2.2, pode-se observar que o histograma lembra a distribuição exponencial. O parâmetro μ pode ser ser estimado através da média amostral, $\bar{X} = 0,10140$. Pode-se calcular, por exemplo, a frequência esperada entre 0 e 0.05 fazendo-se $\int_0^{0,05} 1/0,10140 e^{-x/0,10140} dx = 0,3892$. Desta forma, para os 100 tempos medidos no exemplo espera-se que 38,92 observações estejam neste intervalo. A lista completa é apresentada a seguir:*

Intervalo	f_o	f_e	*D*
$0,00 \leq x < 0,05$	41	$38,92410$	$0,11071$
$0,05 \leq x < 0,10$	19	$24,2908$	$1,15240$
$0,10 \leq x < 0,15$	21	$14,21913$	$3,23368$
$0,15 \leq x < 0,20$	8	$8,38601$	$0,01777$
$0,20 \leq x < 0,50$	10	$11,17893$	$0,12433$

No exemplo, calcula-se $D = 4,6389$.

Para uma aderência perfeita, D deveria ser zero. Contudo, devido à aleatoriedade, irá existir um valor residual para D - pode ser mostrado que D possui distribuição de Chi-Quadrado com $k-1$ graus de liberdade.

O teste de aderência de Chi-quadrado compara o valor calculado de D com os valores críticos para distribuição de Chi-quadrado, apresentada no Apêndice A, com parâmetros $\chi_{1-\alpha,k-r-1}$, onde α é o nível de significância e k é o número de graus de liberdade (é o número de classes) e r é o número de estimadores da distribuição em estudo. No caso da distribuição exponencial, $r = 1$.

A partir deste resultado, consideram-se as seguintes hipóteses:

H_0 A variável aleatória em estudo pode ser modelada com a distribuição teórica sob hipótese com os parâmetros estimados;

H_1 A variável aleatória em estudo não pode ser modelada com a distribuição teórica sob hipótese com os parâmetros estimados.

Para que a hipótese H_0 não seja rejeitada, $D < \chi_{1-\alpha,k-r-1}$

Exemplo 2.4 *Continuando o exemplo anterior, obtemos:*

- *O número de classes é 5 ($k = 5$);*
- *O número de estimadores da distribuição em estudo é $r = 1$;*

Utilizando o nível de significância de $\alpha = 0,1$, obtemos $\chi_{0,9;3} = 6,251$. Logo, a hipótese da aderência à distribuição exponencial não pode ser rejeitada ao nível de significância de 0,1 ($D < \chi_{(1-\alpha,k-r-1)}$: $4,6389 < 6,251$).

2.6.4 Teste de Aderência Kolmogorov-Smirnov

O teste de aderência Kolmogorov-Smirnov (KS) é baseado na observação que a diferença entre a distribuição acumulada observada, $F_o(x_i)$, e a esperada, $F_e(x_i)$, deve ser pequena.

Para que a estatística seja calculada, os dados empíricos $x_1, x_2, \ldots, x_N$ devem ser *ordenados*. N representa a quantidade de amostras disponívies, enquanto i representa o número de observações e $F_o(x_i) = i/N$.

Suponha novamente as hipóteses:

H_0 A variável aleatória em estudo pode ser modelada com a distribuição teórica sob hipótese com os parâmetros estimados;

H_1 A variável aleatória em estudo não pode ser modelada com a distribuição teórica sob hipótese com os parâmetros estimados.

Para realizar o teste, deve ser calculada a seguinte estatística

$$D = max_{1 \leq i \leq N} \left(\left| \frac{i}{N} - F_e(x_i) \right|, \left| F_e(x_i) - \frac{i-1}{N} \right| \right) \tag{2.14}$$

Onde D indica o desvio máximo entre distribuição acumulada observada e a esperada (ou teórica). Para que o a hipótese H_0 não seja rejeitada, o valor de D deve ser menor que o valor crítico para o teste, apresentado pela Tabela B, com um nível de significância α e o número de amostras N.

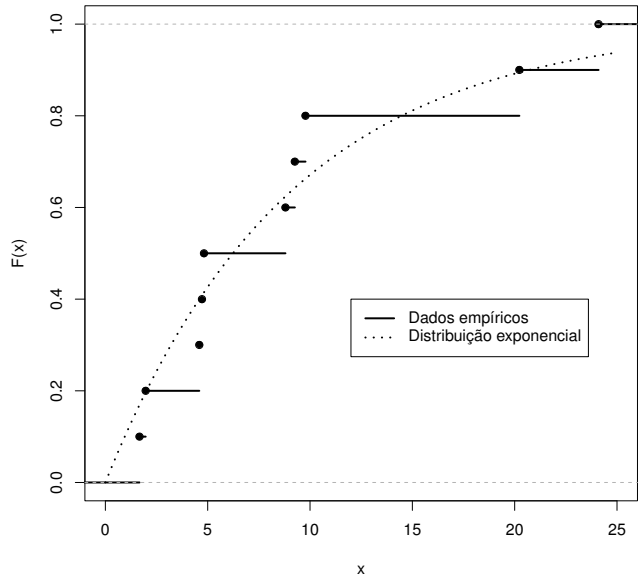


Figura 2.9: $F_o(x_i)$ e $F_e(x)$ para o teste de aderência de Kolgomorov-Smirnov

Exemplo 1: Suponha que a observação de $N = 10$ valores para uma variável aleatória apresentou os seguintes resultados: 8,8; 4,72; 1,97; 9,78; 20,23; 9,26; 24,1; 4,82; 1,67 e 4,59. A Figura 2.9 mostra o $F_o(x_i)$ para dados empíricos, apresentados na figura com uma linha contínua, comparados com a distribuição teórica exponencial, com frequência esperada dada por $F_e(x_i)$).

Para o exemplo, deve-se calcular:

i	x_i	$\frac{i-1}{N}$	$\frac{i}{N}$	$F_e(x_i)$	D
1	1,67	0,00	0,10	0,17	0,17
2	1,97	0,10	0,20	0,20	0,01
3	4,59	0,20	0,30	0,40	**0,20**
4	4,72	0,30	0,40	0,41	0,11
5	4,82	0,40	0,50	0,41	0,09
6	8,80	0,50	0,60	0,62	0,124
7	9,26	0,60	0,70	0,64	0,06
8	9,78	0,70	0,80	0,66	0,14
9	20,23	0,80	0,90	0,89	0,10
10	24,10	0,90	1,00	0,93	0,07

e encontramos o valor de $D = 0,20$.

Para rejeitar H_0, o resultado deve ser comparado com os valores críticos da estatística KS, dado por $K_{\alpha,n}$, onde n é o número de observações e α é o nível de significância desejado, conforme apresentado no Apêndice B. Neste exemplo, considerando um nível de significancia de $\alpha = 0,1$, com o valor de $N = 10$, o valor crítico é de 0.4566, de forma que a hipótese H_0 não pode ser rejeitada.

□

O teste KS é melhor do que o teste de Chi-quadrado quando o tamanho da amostra é pequeno. Em situações em que o tamanho da amostra é grande, os testes são equivalentes. O KS pode ser aplicado com distribuições contínuas e com distribuições discretas.

2.6.5 QQPlot

O QQPlot é uma ferramenta gráfica utilizada para comparar características de duas populações. Nesta técnica, dadas duas distribuições acumuladas F e G, associadas a funções quantile F^{-1} e G^{-1} (a inversa da distribuição acumulada é chamada função quantile). Suponha que os quantiles de uma das funções seja dado por $q_1, q_2, ..., q_n$, o QQPlot é criado plotando as coordenadas $(F(q_i), G(q_i), 1 \leq i \leq n)$.

Desta forma, o principal problema é encontrar o número de quantiles a serem utilizados. Quantiles são pontos tomados em intervalos regulares na distribuição acumulada de probabilidade de uma certa variável aleatória. Supondo que o número de quantiles seja dado por n, teremos n sub conjuntos de dados - as fronteiras são denominados quantiles.

Nesta técnica, o conjunto de dados é ordenado em ordem de grandeza, sendo os valores que dividem o conjunto em quatro partes iguais chamados quartis, em dez partes os decis, e em N partes, que podem corresponder ao número de dados do conjunto, os quantis. Neste gráfico, os

pontos representam os quantis de cada uma das amostras, colocados nos eixos x e y. Se as duas amostras vêm da mesma população, os pontos devem estar em torno da linha diagonal em 45^o sobre a origem. Comparando-se os pontos traçados no gráfico com esta linha diagonal, caso os pontos estejam em uma linha paralela à diagonal, as duas populações possuem distribuição semelhante e um processo está localizado em um nível mais alto em relação ao outro.

O QQPlot pode ser utilizado para comparar duas populações quaisquer, inclusive com tamanhos diferentes; no entanto, para realização de um teste de aderência, uma das populações pode ser obtida a partir da distribuição teórica em hipótese.

As duas populações possuem funções de probabilidade $F^-1(x)$ e $G^-(x)1$, que representam a inversa da distribuição acumulada das duas distribuições sendo comparadas. O QQPlot plota o enésimo quantile de F contra o enésimo quantile de G.

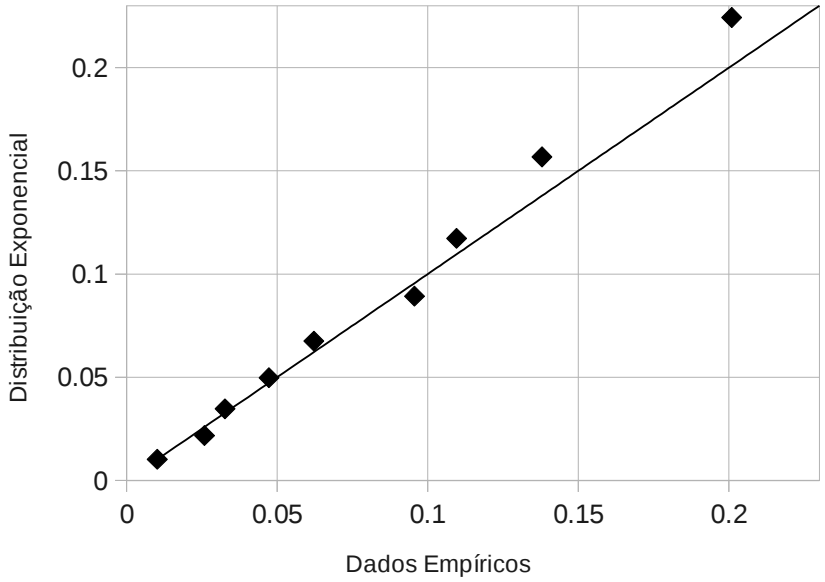
Exemplo 1: Suponha os intervalos de tempo entre requisições, apresentados na Tabela 2.2. Para montagem dos quantiles, vamos considerar dividir o conjunto de dados em $n = 10$ intervalos (neste caso, o quantile é chamado *decile*). Ordenando os dados da Tabela 2.2, obtemos:

0,0018	0,0102	0,0271	0,0337	0,0479	0,0624	0,1023	0,1101	0,1448	0,2188
0,0020	0,0105	0,0273	0,0345	0,0518	0,0673	0,1024	0,1169	0,1557	0,2188
0,0024	0,0109	0,0277	0,0357	0,0532	0,0699	0,1027	0,1193	0,1605	0,2607
0,0059	0,0112	0,0281	0,0358	0,0579	0,0716	0,1028	0,1219	0,1622	0,2935
0,0066	0,0138	0,0285	0,0366	0,0579	0,0719	0,1031	0,1221	0,1628	0,2957
0,0074	0,0141	0,0286	0,0396	0,0584	0,0754	0,1058	0,1226	0,1700	0,3216
0,0089	0,0166	0,0306	0,0403	0,0595	0,0788	0,1088	0,1246	0,1712	0,3773
0,0092	0,0187	0,0308	0,0434	0,0603	0,0823	0,1090	0,1291	0,1952	0,4676
0,0095	0,0237	0,0325	0,0454	0,0613	0,0943	0,1092	0,1308	0,1983	0,4973
0,0102	**0,0259**	**0,0327**	**0,0473**	**0,0622**	**0,0956**	**0,1096**	**0,1379**	**0,2009**	**0,5314**

Tomando-se os dados ordenados desta forma, com 100 amostras para obter os limites para obter o aumento na probabilidade acumulada de $0,1$ dividimos os dados em 10 colunas (100/10), onde é possível visualizar os valores limite para x que divide os dados em $n = 10$ porções iguais, que estão marcados em negrito na tabela acima, representando diversos quantiles. Neste caso, a hipótese H_0 é a aderência da distribuição exponencial, com média $\mu = 0.097409$. Para obter o primeiro quantile, fazemos $F(x) = 0,1$, utilizando a distribuição exponencial $F(x) = 1 - e^{-x/\mu}$, de onde se obtem o valor do primeiro quantile em $x = 0,0103$. Completando os cálculos, obtemos:

i	$P(X \leq x)$	Quantile empírico	Quantile teórico (dist. exp.)
1	0,1	0,0102	0,0103
2	0,2	0,0259	0,0217
3	0,3	0,0327	0,0347
4	0,4	0,0473	0,0498
5	0,5	0,0622	0,0675
6	0,6	0,0956	0,0893
7	0,7	0,1096	0,1173
8	0,8	0,1379	0,1568
9	0,9	0,2009	0,2243
10	1,0	0,5314	

Utilizando-se os dados obtidos, é possível plotar o gráfico a seguir:



Observe que se as duas populações fossem provenientes da mesma distribuição, os pontos no gráfico estariam exatamente sobre a linha diagonal. Neste caso, observamos que os quantiles dos dados empíricos diferem pouco dos quantiles da distribuição exponencial, o que permite supor que a distribuição das duas variáveis é semelhante. □

No exemplo, os quantiles foram estabelecidos para uma função empírica e também para a distribuição exponencial. Estes casos são relativamente simples - no entanto, em muitos casos, existe grande dificuldade para obtenção dos quantiles. Este problema é tratado por vários autores.

O QQPlot é uma das ferramentas mais poderosas para verificação da aderência de duas populações, apesar de não ser um método que permita uma medida quantitativa da qualidade da aderência entre as duas variáveis.

2.7 Exercícios

Exercício 1: Suponha uma variável aleatória X com distribuição triangular, cuja função densidade de probabilidade é apresentada na Tabela 2.1, com $a = 1, m = 2$ e $b = 4$. Determine:

1. Função densidade de probablidade $f(x)$;
2. Função distribuição acumulada de probabilidade $F(x)$;
3. $P(X \leq 2)$;
4. $P(1 \leq X \leq 3)$;

□

Exercício 2: Realize o teste de aderência para os dados a seguir, que representam o tamanho do pacote transmitido pelo servidor.

0.76	5.08	16.02	5.76	1.36	4.83	15.89	5.85	11.23	11.16
2.85	12.54	0.48	51.40	6.72	16.50	1.71	2.25	7.77	4.84
4.65	17.15	23.51	1.91	1.11	3.20	9.13	2.36	19.93	8.15
15.35	27.60	31.11	8.75	4.69	0.86	33.85	1.87	7.45	10.24
18.50	14.85	4.39	22.05	14.60	18.62	11.96	25.90	0.47	2.72
3.69	5.41	0.60	2.40	8.96	1.79	6.67	2.11	4.68	4.32
13.77	1.72	3.39	12.15	13.52	10.85	15.39	12.29	17.87	3.72
4.28	11.25	0.88	12.17	8.33	19.10	7.95	3.05	15.00	9.02
6.95	0.89	0.12	8.19	3.55	9.27	12.83	15.41	15.83	1.02
1.47	7.82	2.94	5.52	2.82	55.41	0.90	0.32	12.16	25.10

Você deve apresentar:

1. Teste Kolmogorov-Smirnov;
2. Teste de Chi-Quadrado;
3. QQPlot.

□

Exercício 3: Realize o teste de aderência para os dados a seguir, que representam o tamanho do pacote transmitido pelo servidor.

10.12	15.54	-0.48	6.96	15.91	15.32	2.47	16.50	16.31	10.48
10.92	5.80	8.62	14.91	14.27	18.42	8.72	10.52	11.80	7.80
8.77	7.82	5.94	17.77	8.97	10.93	3.95	4.28	13.32	7.45
6.18	6.72	5.26	6.47	10.35	1.91	3.50	11.18	18.53	22.31
4.42	10.37	5.68	10.30	8.29	23.78	9.72	13.62	10.48	4.08
3.46	17.82	9.37	7.43	15.90	15.59	15.80	22.70	22.83	11.58
7.37	6.09	15.88	9.64	9.47	8.83	10.30	10.68	-4.86	9.75
10.60	4.45	-1.72	13.23	12.02	8.42	13.19	4.48	4.63	7.03
6.67	6.41	6.10	12.66	10.50	3.55	10.90	8.96	3.39	10.83
1.33	8.64	13.92	17.08	15.49	6.48	7.79	8.58	14.12	7.81

Você deve apresentar:

1. Teste Kolmogorov-Smirnov;
2. Teste de Chi-Quadrado;
3. QQPlot.

□

Exercício 4: Fazer uma coleta de dados em um sistema. Escolha uma das seguintes variáveis:

1. Tamanho de arquivos no Unix;
2. Tamanho de arquivos no Windows;
3. Tamanho de arquivos em um servidor de páginas;
4. Outras ideias são bem vindas.

Você deve realizar a caracterização de uma destas variáveis utilizando distribuições de probabilidade: (a) identificar a distribuição teórica que melhor se adapta a VA observada e (b) realizar os testes de aderência de Chi-Quadrado, Kolgomorov-Smirnov e também o QQPlot para confirmar sua hipótese. □