Modelagem e Avaliação de Desempenho

Prof. Carlos Marcelo Pedroso

2025

Análise de Desempenho: Abordagens

- **Modelos Matemáticos/Analíticos:** fórmulas e equações matemáticas para prever o comportamento do sistema.
  - ✔ Geralmente, é a abordagem *mais rápida e de menor custo*.
  - ✔ Permite obter *insights profundos* sobre as relações de causa e efeito.
  - ✔ Pode ser *simplificado demais* para sistemas muito complexos ou com comportamentos altamente dinâmicos e imprevisíveis.
  - ✔ *Exemplos:*
    - ✓ Modelos de *teoria das filas* (M/M/1, M/G/1, etc.).
    - ✓ Cadeias de Markov, Modelo Ocultos de Markov (HMM).
    - ✓ Séries Temporais (ARIMA, FARIMA).
    - ✓ Redes de Petri.
    - ✓ Network Calculus.
- A escolha da abordagem depende da natureza do problema, dos dados disponíveis e da precisão desejada.

Análise de Desempenho: Abordagens

- **Modelagem e Simulação:** *modelo computacional* que busca replicar o comportamento do sistema real.
  - ✔ Lida bem com a *complexidade, aleatoriedade e interações* do sistema.
  - ✔ Permite testar cenários hipotéticos sem impactar o sistema real.
  - ✔ Pode ser *custoso em tempo e recursos computacionais*, e a validação do modelo é crucial. A generalização de resultados pode ser difícil.
  - ✔ *Principais simuladores utilizados:*
    - ✓ NS-3 (Network Simulator 3): Simulador mais utilizado e respeitado na comunidade acadêmica e de pesquisa.
    - ✓ OMNeT++: É popular em pesquisas sobre redes ad-hoc, redes de sensores, redes veiculares (VANETs) e sistemas distribuídos.
    - ✓ OPNET/Riverbed Modeler (proprietário): É usado em ambientes corporativos e acadêmicos que podem bancar a licença, para estudo de desempenho de redes empresariais, provedores de serviço e redes sem fio.
    - ✓ Cisco Packet Tracer: é uma ferramenta de simulação e visualização de rede desenvolvida pela Cisco Systems. Permite a criação, configuração e testes de topologias de rede em um ambiente virtual. O objetivo é didático.

Exemplo: Análise de Desempenho de um Servidor Web

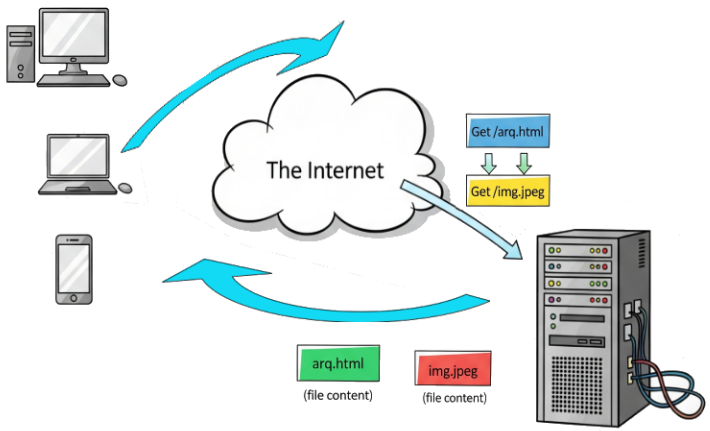
- Suponha um servidor Web que processa requisições e envia arquivos como resposta.
  - Considere que quantidade de dados transferidos (b, em bits) é o produto da taxa de transmissão (r, em bps) pelo tempo (t em segundos) gasto:

$$b = r \times t$$

  - Esta equação é um modelo determinístico. A taxa de transmissão r de placa de rede do servidor é constante, mas existem variações no:
    - Taxa de chegada de requisições.
    - Tamanho dos arquivos solicitados.
    - Congestionamento da rede.
    - Carga no próprio servidor.
  - Diante dessa variabilidade, que tipo de estatísticas seriam relevantes para medir e caracterizar o comportamento desse servidor Web ao longo do tempo?

Exemplo: Análise de Desempenho de um Servidor Web

- Considere que foram coletadas as médias e desvio padrão para as seguintes variáveis:
  - ✔ Taxa de chegada de requisições. (λ).
  - ✔ Tamanho dos arquivos solicitados (τ).
- Considere que a taxa de transmissão r da placa de rede é constante.
- Estas medidas são suficientes para planejar/prever o desempenho do sistema? Considere um exemplo:



- ✔ $\overline{\lambda}$=10 req/s;
- ✔ $\overline{\tau}$= 10000 bits;
- ✔ Considere r constante.

Estocástico

- A palavra estocástico significa "aleatório", "chance". Seu antônimo é "determinístico", "certo".
- Um modelo estocástico prevê um conjunto de possibilidades de acontecimentos.
- Exemplo: um lançamento de moeda, com possibilidade $\frac{1}{2}$ para cada um dos resultados.

Frequência Relativa

- Suponha que repetimos n vezes o experimento $\mathcal{E}$, e sejam A e B dois eventos associados a $\mathcal{E}$.
- Admita que sejam n_A e n_B o número de vezes que o evento A e o evento B ocorrem.
- $f_A = n_A/n$ é denominada frequência relativa do evento A nas n repetições de $\mathcal{E}$;
- É claro que $0 \leq f_A \leq 1$;
- ... e $n = n_A + n_B$.

Probabilidade

- Seja $\mathcal{E}$ um experimento.
  - ✔ Seja S um espaço amostral associado a $\mathcal{E}$.
  - ✔ Por exemplo, se $\mathcal{E}$ representa o lançamento de uma moeda, o espaço amostral é $S = \{\text{cara}, \text{coroa}\}$.
  - ✔ Por exemplo, se $\mathcal{E}$ representa uma nota na UFPR, o espaço amostral é $S : \{0, 1, 2, ..., 99, 100\}$.
- A cada evento A associaremos um número real representado por $P(A)$ e denominado probabilidade de A, que satisfaça as seguintes propriedades:
  - ✔ $0 \leq P(A) \leq 1$;
  - ✔ $P(S) = 1$;

Variável Aleatória

- Sejam $\mathcal{E}$ um experimento e S um espaço amostral associado ao experimento.
- Uma função X, que associe a cada elemento de S um número real, $X(s)$, é denominada variável aleatória.
- Note que s pode assumir qualquer valor pertencente ao espaço amostral S de $\mathcal{E}$.
- *Exemplo: Copa do Mundo*. Considere o número de gols marcados por cada time em uma partida é uma VA.
  - ✔ Qual o espaço amostral?
  - ✔ Esta VA é contínua ou discreta?

Variáveis Aleatórias - Notação

- Letras maiúsculas, como X, Y ou Z, denotam variáveis aleatórias;
- Letras minúsculas, como x, y ou z, denotam resultados possíveis pertencentes ao espaço amostral;
- A expressão $\{X = x\}$ é o evento que a variável aleatória assume um valor igual a x.
- A expressão $\{X \leq x\}$ é o evento que a variável aleatória assume um valor menor ou igual a x.

Variável Aleatória: Exemplos

- Suponha que foi medido o tráfego em um servidor de páginas. São exemplos de variáveis aleatórias:
  - ✔ V = índice de inflação mensal no Brasil;
  - ✔ X = duração de uma ligação telefônica;
  - ✔ Y = intervalo entre pacotes em uma rede;
  - ✔ W = tempo de atendimento da requisição em um servidor Web.
  - ✔ Z = número de requisições recebidas a cada segundo por um servidor Web;

... anotar o espaço amostral para cada caso...

Função Densidade de Probabilidade

- $f(x)$ é a função que indica a probabilidade de obter-se exatamente o valor x em um experimento, ou seja, $P(X = x)$;
- Normalmente, a função densidade de probabilidade é denotada por letras minúsculas: $f(x) = P(X = x)$;

Anotar: qual a relação necessária para que $f(x)$ represente uma distribuição de probabilidade?

Função Densidade de Probabilidade

- Desta forma, pode-se escrever:
  - ✔ Para variávies aleatórias discretas:

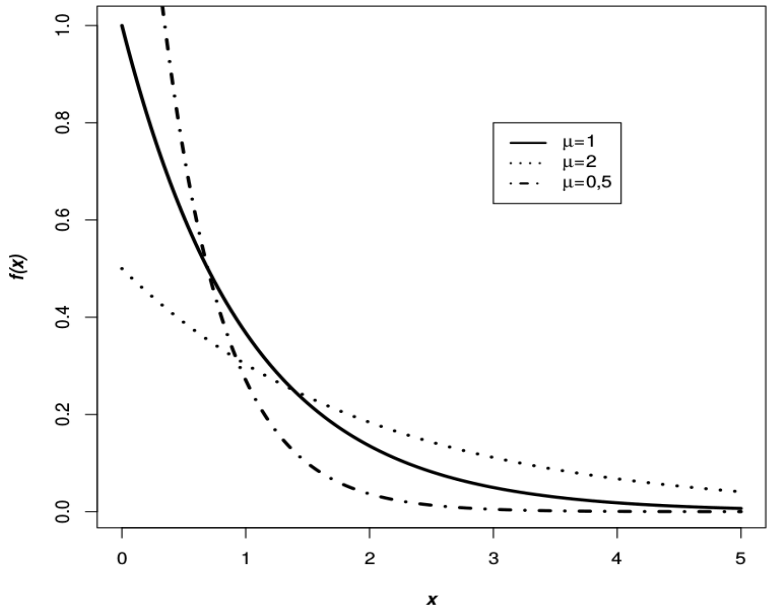
$$P\{a \leq X \leq b\} = \sum_{i=a}^{b} f(x_i)$$

  - ✔ Para variáveis aleatórias contínuas:

$$P\{a < X \leq b\} = \int_{a}^{b} f(x)dx$$

Exemplo: Distribuição Exponencial

Parâmetros	μ : 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
Variância ($Var[X]$)	μ^2



Função Distribuição Acumulada

- A função de distribuição acumulada de probabilidade é definida da seguinte forma:
  - ✔ $F(x) = P(X \leq x)$
  - ✔ Esta é a distribuição de probabilidade acumulada da variável aleatória X, normalmente denotada por letras capitais: $F(x)$.
- Caso X seja uma variável aleatória contínua, $F(x)$ e $f(x)$ são relacionados por:

$$F(x) = \int_{-\infty}^{x} f(\xi) d\xi, \quad -\infty < x < \infty$$

- Se $F(x)$ é diferenciável em x, então pode-se escrever que X tem uma função densidade de probabilidade dada por:

$$f(x) = \frac{d}{dx} F(x) = F'(x)$$

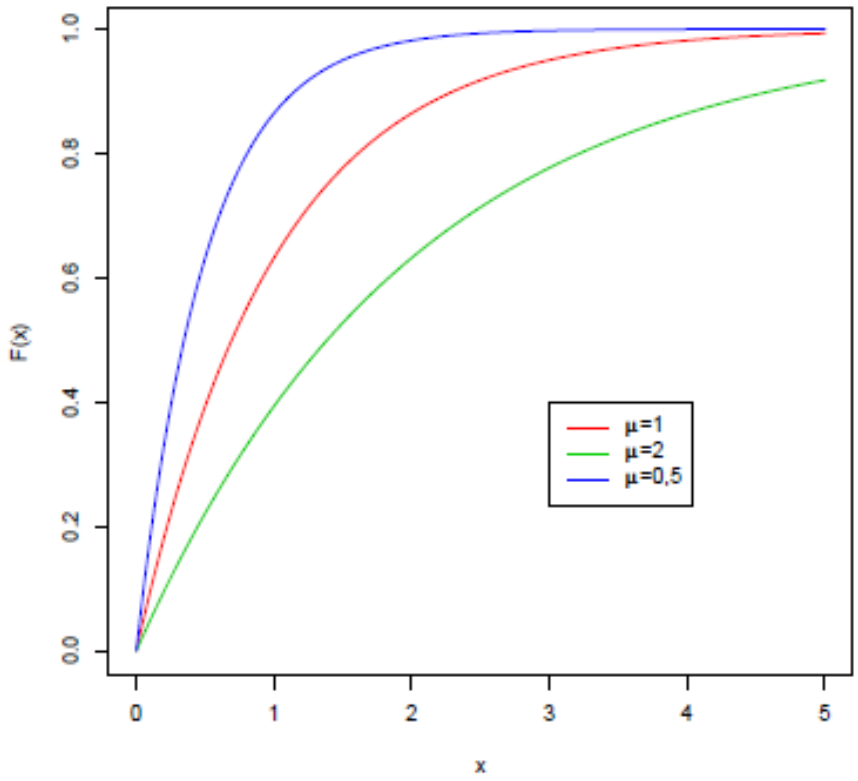
anote o caso da VA discreta

Exemplo

- Esboce qual seria a distribuição de probabilidade do tempo de duração de uma ligação telefônica?
- Considere que o usuário é tarifado por tempo de ligação.
- Qual seria a alteração no formato da curva caso a tarifação seja por valor fixo mensal, independentemente do tempo de duração ou quantidade de chamadas?

Distribuição Exponencial

- A função de distribuição acumulada de probabilidade exponencial é dada por $F(x) = 1 - e^{-\lambda x}$, $x > 0$, onde λ é 1/média.
- Prove como exercício
- Graficamente:



Valor Esperado e Momentos

- Se X é uma V.A. discreta, então o n-ésimo momento em torno de zero é dado por

$$E[X^n] = \sum_i x_i^n P\{X = x_i\}$$

- Quando a soma não converge, o momento não existe.
- Se X é contínuo com função densidade de probabilidade dada por $f(x)$, então o n-ésimo momento em torno de zero é dado por

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

- O primeiro momento central em torno de zero é a esperança $E[X]$: ou seja, a esperança pode ser obtida através de

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

Valor Esperado e Momentos

- O n-ésimo momento central em torno de $E[X]$ é dado por

$$E[(X - E[X])^n]$$

- O primeiro momento central em torno da média é zero.
- O segundo momento central em torno da média é chamado de variância de X, denotado por $V[X]$:

$$V[X] = E[(X - E[X])^2]$$

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

Valor Esperado e Momentos

- A variância é dada por σ^2 ou $Var[X]$.
- Formas equivalentes para variância:
  - ✔ $Var[X] = E[(X - E[X])^2]$
  - ✔ $Var[X] = E[X^2] - E[X]^2$ (obtido facilmente a partir da equação acima)
- A raiz quadrada da variância é chamada de desvio padrão, e denotada por σ.

Exemplo – Distribuição Exponencial - anotar

Exercícios

Considere a distribuição:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x^3 & 0 < x < 1 \\ 1 & x \geq 0 \end{cases}$$

- **Responda:**
  A. Esta realmente uma distribuição de probabilidade? Prove matematicamente.
  B. Calcule a função densidade de probabilidade $f(x)$
  C. $P(X > 0,7)$
  D. $P(0,4 < X \leq 0,9)$
  E. Calcule $E[X]$ e $V[X]$

Exercícios

Suponha a distribuição de probabilidade uniforme, dada por $f(x) = 0,01$ para $0 < x < 100$.

- **Responda:**
  A. Calcule a expressão para $F(x)$.
  B. Calcule $P(X < 20)$ utilizando $f(x)$ e $F(x)$.
  C. Calcule $E[X]$ e $V[X]$.

Estimadores

- A média amostral é obtida a partir da repetição do experimento, e normalmente é denotada por:

$$\bar{\mu} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- A variância amostral[1] é dada por:

$$\bar{\sigma}^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{\mu})^2}{n-1}$$

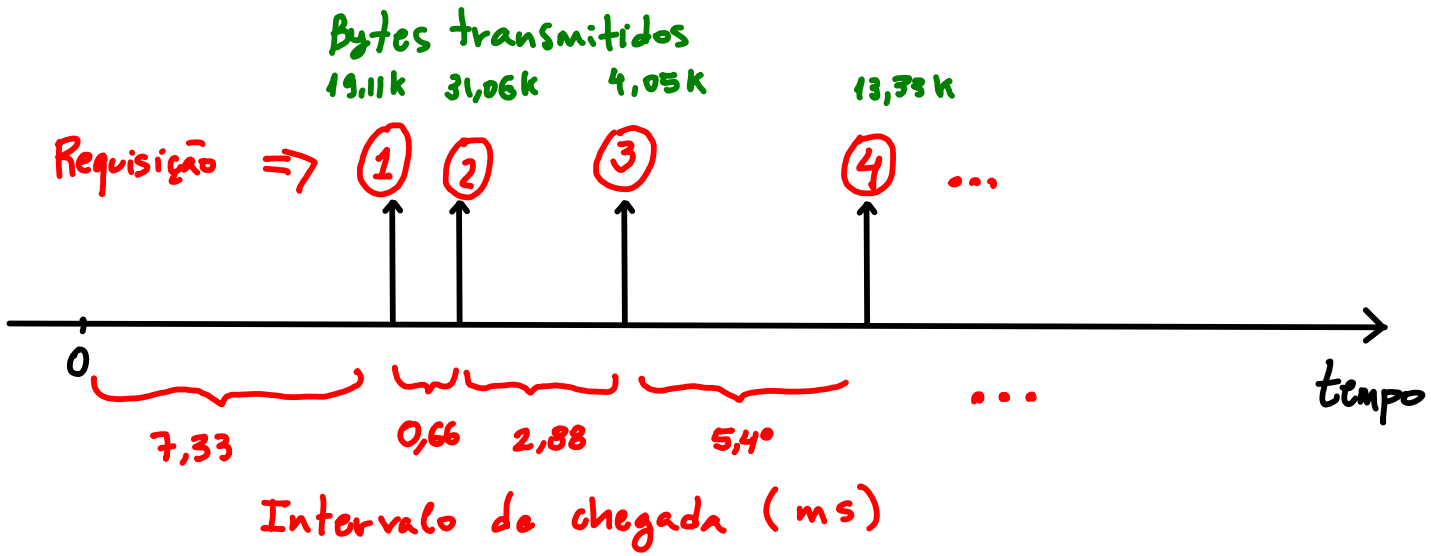
- O desvio padrão amostral é dado pela raiz quadrada da variância. Neste caso, o desvio padrão é dado por $\bar{\sigma}$.

[1] Nota: A fórmula de variância amostral fornecida é a variância amostral. A principal diferença entre variância populacional e variância amostral reside no conjunto de dados utilizado para o cálculo. A variância populacional é calculada usando todos os dados de uma população, enquanto a variância amostral é calculada a partir de um subconjunto dessa população, chamado amostra. A fórmula para calcular a variância amostral utiliza um denominador de $(n-1)$, enquanto a variância populacional usa n, o tamanho total da população.

Exemplo

Suponha que foi realizada a medição das seguintes variáveis em um servidor de páginas:

- Intervalo entre requisições (em milisegundos);
- Quantidade de bytes transmitidos por requisição (em K bytes);



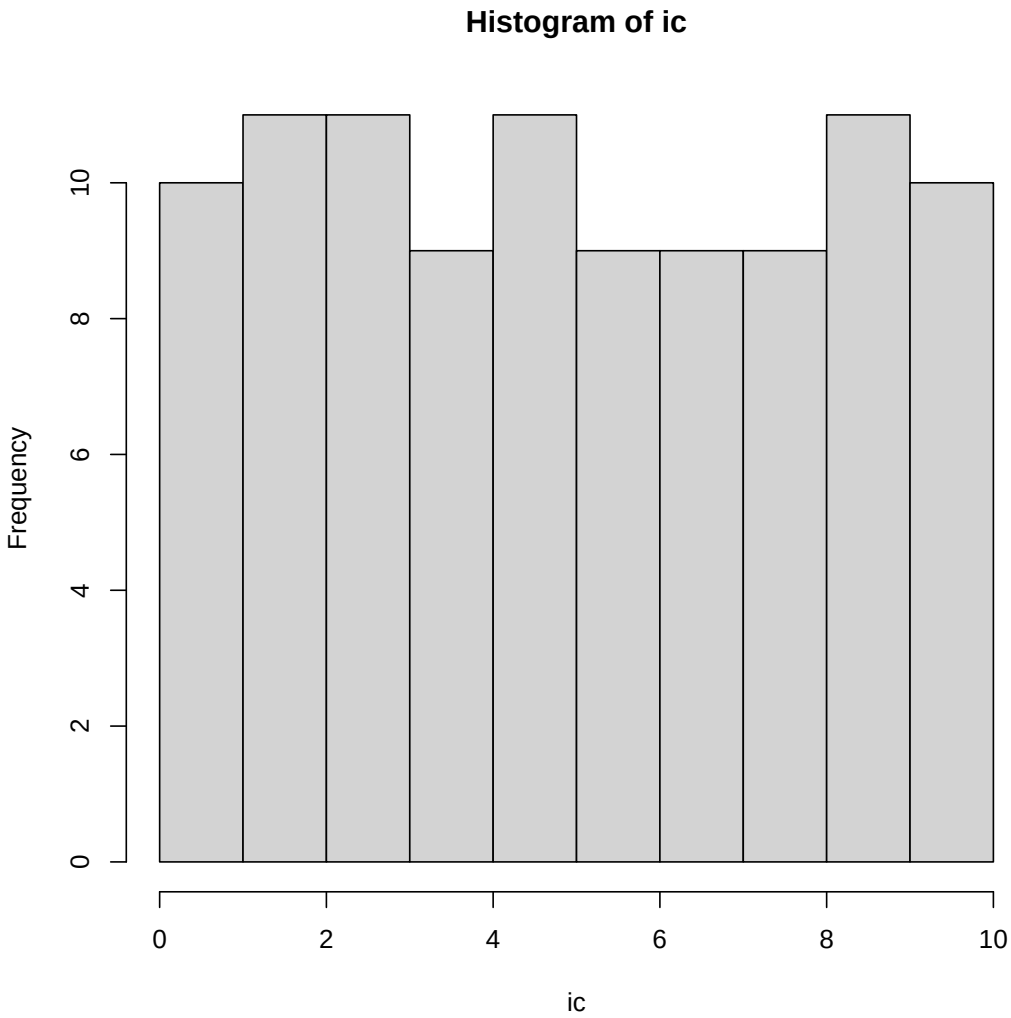
Intervalo entre requisições (em milisegundos)

Dados:

7.33	0.66	2.88	5.48	2.43	3.58	1.06	9.11	8.56	8.47
9.51	1.02	4.74	5.72	1.30	7.18	9.05	4.21	7.86	1.37
6.35	8.84	4.96	6.45	4.87	6.63	2.18	8.58	0.89	4.68
8.61	9.01	4.49	2.11	1.41	5.27	0.00	6.61	3.62	9.22
5.14	4.27	1.87	3.50	7.23	0.64	0.94	1.92	7.67	4.66
7.99	2.89	9.30	9.31	1.28	3.39	5.07	3.97	8.16	1.69
4.43	7.97	8.07	5.67	7.73	3.41	6.90	2.45	4.74	3.62
3.22	2.12	5.25	2.56	1.41	8.53	9.99	0.65	4.90	9.89
3.53	8.45	5.37	9.12	8.25	0.09	8.49	5.96	2.16	0.05
0.51	2.07	6.33	0.60	7.27	6.80	6.58	6.20	2.93	1.09

$$\bar{\mu} = 4.9255, \bar{\sigma} = 2.92761$$

Histograma - intervalo entre requisições



Intervalo	Fo
0 ⊣ 1	10
1 ⊣ 2	11
2 ⊣ 3	11
3 ⊣ 4	9
4 ⊣ 5	11
5 ⊣ 6	9
6 ⊣ 7	9
7 ⊣ 8	9
8 ⊣ 9	11
9 ⊣ 10	10

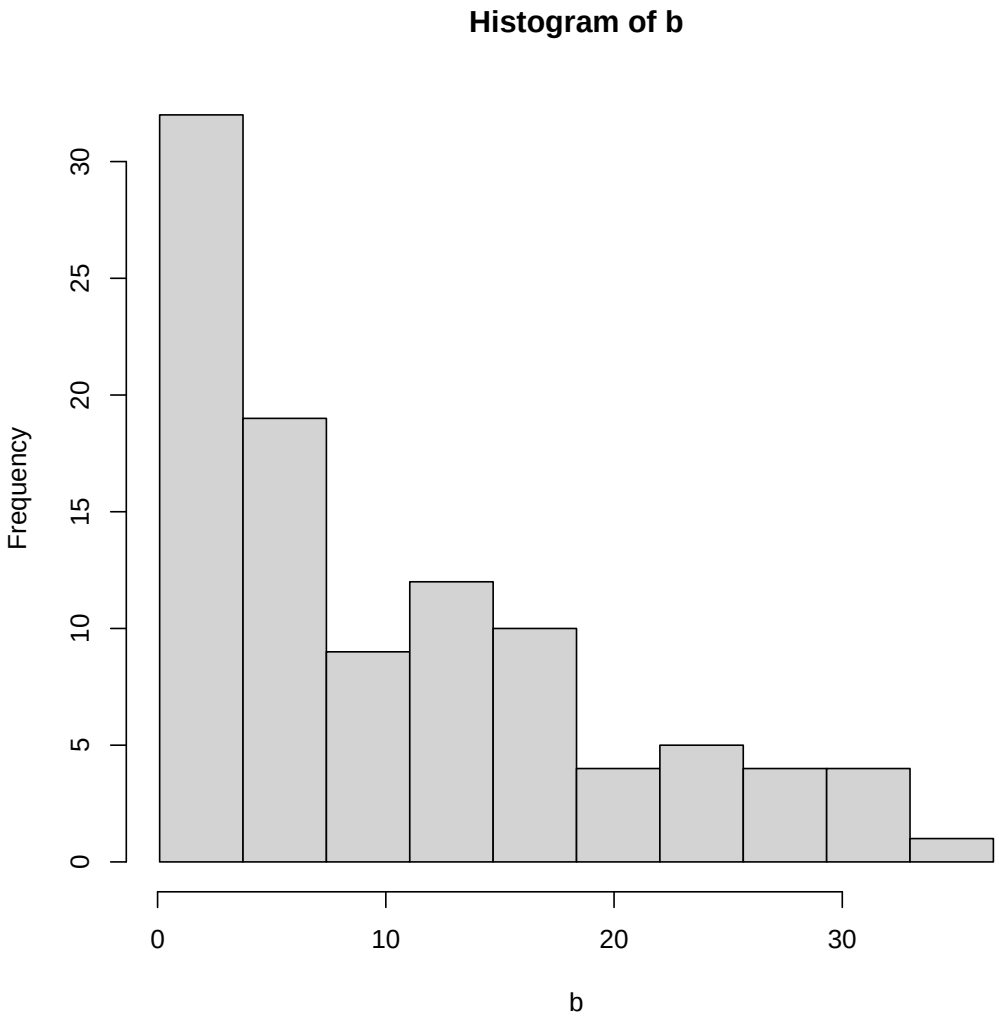
Quantidade de bytes transmitidos por requisição (em K bytes);

Dados:

19.11	31.06	4.05	13.33	1.51	4.05	19.66	5.23	3.54	24.90
26.71	14.59	9.97	27.39	0.22	9.15	9.62	13.51	12.53	16.21
10.60	17.98	15.15	12.20	13.31	3.61	0.35	5.10	22.83	2.13
1.50	0.09	2.26	6.72	15.92	31.01	2.57	16.42	6.29	8.30
13.94	1.55	7.01	21.14	1.34	13.91	5.54	7.38	6.62	5.90
5.95	3.41	13.40	25.12	22.15	12.41	4.91	30.34	9.63	6.35
3.47	31.31	15.95	20.57	26.79	1.37	4.86	3.64	2.33	3.00
6.02	0.72	12.55	0.23	16.78	9.86	25.94	16.52	6.52	3.63
1.48	3.35	10.20	3.48	22.01	0.38	1.02	36.62	7.36	15.97
2.89	17.35	0.96	1.48	1.71	14.62	5.38	2.64	8.62	1.65

$$\bar{\mu} = 10.49, \bar{\sigma} = 8.92$$

Histograma - intervalo entre requisições



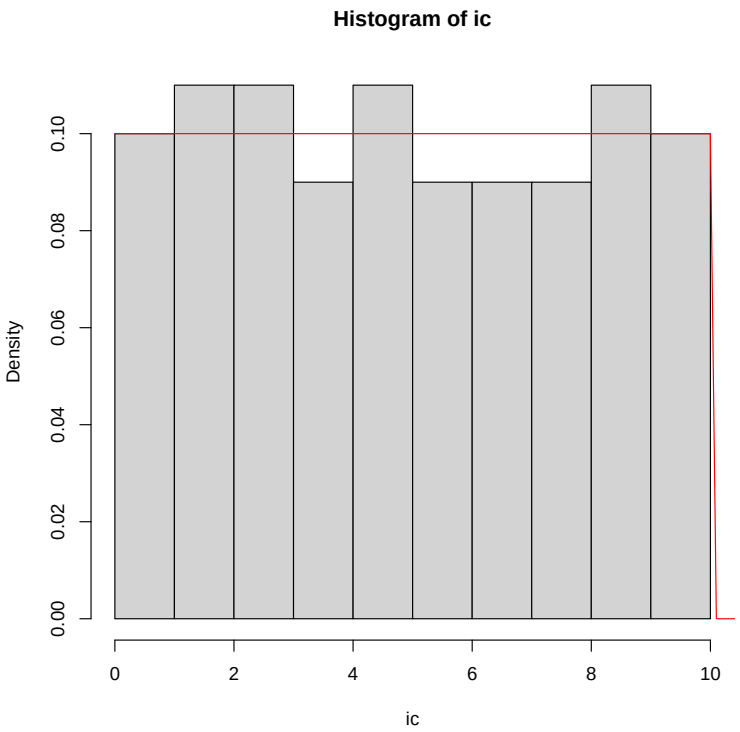
Intervalo	Fo
0.090 ⊣ 3.743	32
3.743 ⊣ 7.396	19
7.396 ⊣ 11.049	9
11.049 ⊣ 14.702	12
14.702 ⊣ 18.355	10
18.355 ⊣ 22.008	4
22.008 ⊣ 25.661	5
25.661 ⊣ 29.314	4
29.314 ⊣ 32.967	4
32.967 ⊣ 36.620	1

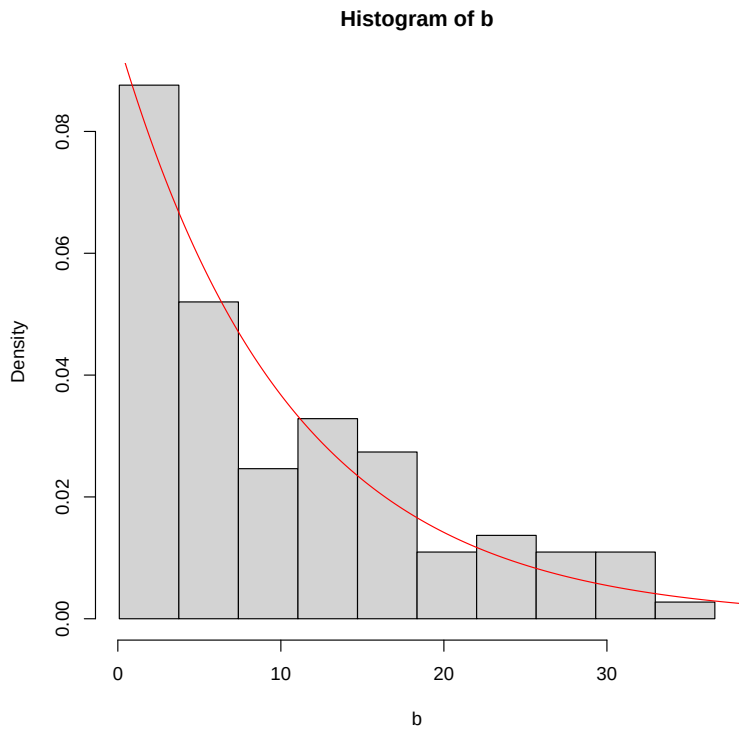
Comandos R

- Comandos com o R:

```
> b<-scan("b1.dat")
> limites_classes <- seq(min(b), max(b), length.out = 11)
> limites_classes
  [1]  0.090  3.743  7.396 11.049 14.702 18.355 22.008 25.661 29.314 32.967
  [11] 36.620
> classes_dados <- cut(b, breaks = limites_classes, include.lowest = TRUE)
> contagem_por_classe <- table(classes_dados)
> contagem_por_classe
  [0.09,3.74]  (3.74,7.4]    (7.4,11]   (11,14.7] (14.7,18.4]   (18.4,22]
           32           19           9          12          10           4
  (22,25.7] (25.7,29.3]   (29.3,33]   (33,36.6]
            5           4           4           1
> hist(b, breaks = limites_classes)
# Comparação com a distribuição exponencial
> hist(b, breaks = limites_classes,freq=FALSE)
> x<-seq(0,100,0.1)
> lines(x,dexp(x,rate=1/mean(b)),col="red")
```

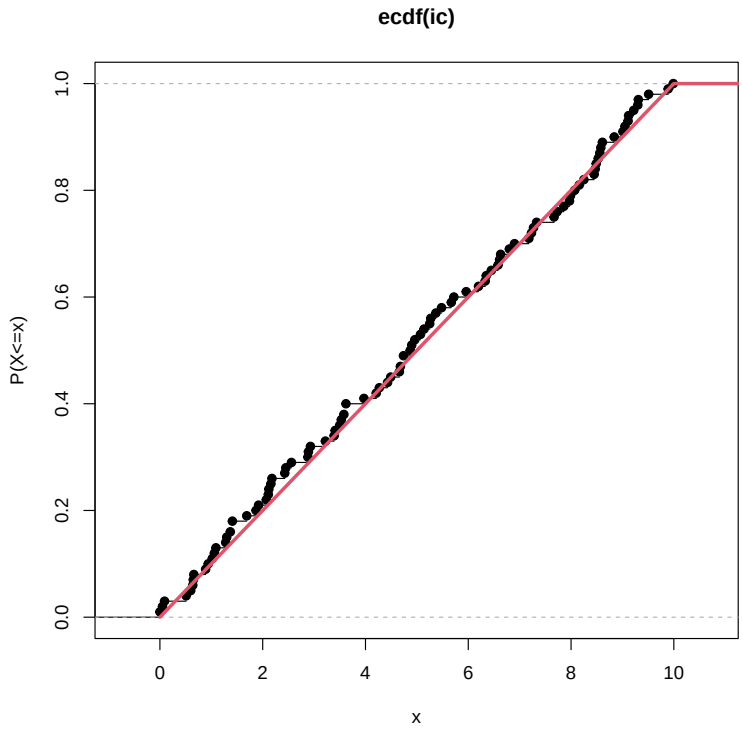
Teste de Aderência

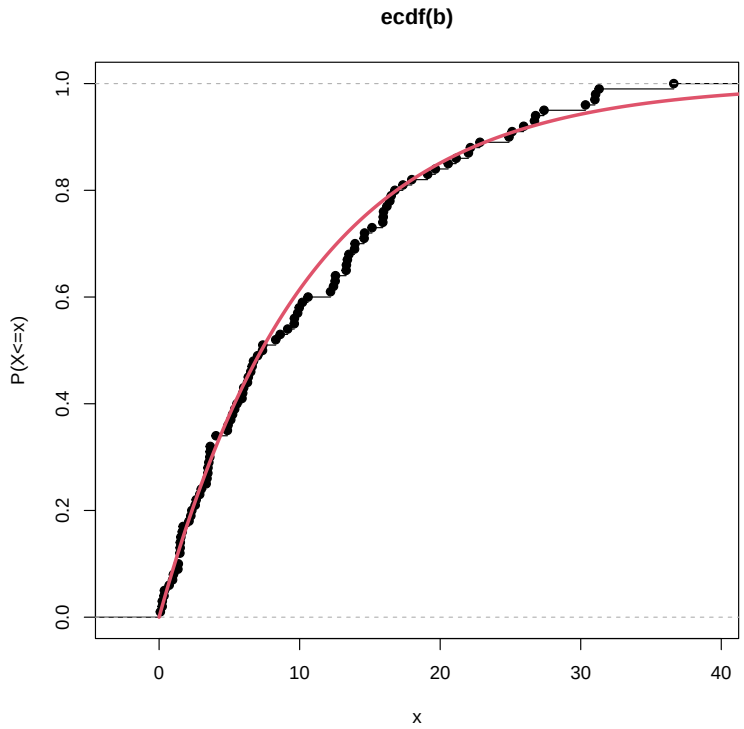




- Os histogramas possuem semelhanças com as distruições Uniforme e Exponencial, respectivamente;
- As linhas em vermelho nos histogramas acima mostram a distribuição teórica.

Teste de Aderência





- Podemos fazer a comparação também utilizando a distribuibuição acumulada, como pode ser visualizado nas figuras acima, onde foi plotada em preto a distribuição empírica acumulada. As linhas em vermelho mostram a distribuição teórica.
- Existem diversos testes de a aderência disponívies: Chi-Quadrado, Kolmogorov-Smirnov (K-S), Anderson-Darling, Shapiro-Wilk, QQPlot.

Comandos R

- Para a comparar a distribuição acumulada:

```
> x<-seq(0,100,0.1)
> pdf("ecdf1.pdf")
> plot(ecdf(ic),ylab="P(X<=x)")
> lines(x, punif(x,min=0,max=10),col=2, lwd = 3)
> dev.off()
> pdf("ecdf2.pdf")
> plot(ecdf(b),ylab="P(X<=x)")
> lines(x, pexp(x,rate=1/mean(b)),col=2, lwd = 3)
> dev.off()
```

- Eu prefiro a abordagem comparando com a distribuição acumulada.

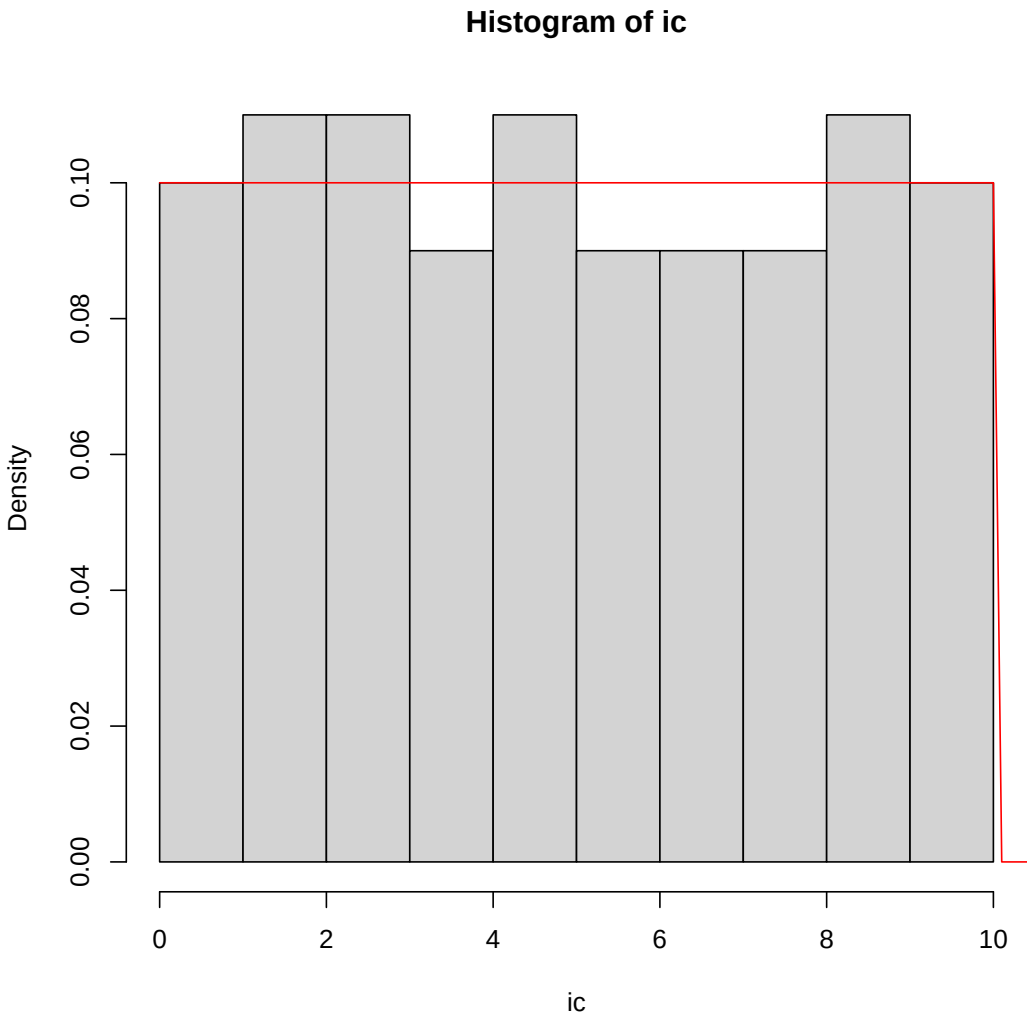
Teste de Aderência de Chi-Quadrado

- Os histogramas possuem semelhanças com as distruições Uniforme e Exponencial, respectivamente;
- O teste de Chi-Quadrado é um teste de aderência, em que o resultado é **rejeitar ou não** a hipótese da aderência dos dados amostrais à uma determinada distribuição com um nível de significância desejado.
- Calcula-se:

$$D = (f_o - f_e)^2 / f_e;$$

- f_e é a frequência esperada (obtida a partir da distribuição teórica e f_o é a frequência observada;
- O teste de aderência de Chi-quadrado compara o valor de D com o valor tabelado da distribuição de Chi-quadrado:
  - ✔ $\chi^2_{(1-\alpha,k-r-1)}$
  - ✔ onde α é o nível de significância e k é o número de classes e r é o número de estimadores da distribuição em estudo.
  - ✔ Para que a hipótese seja aceita, D deve ser menor que $\chi^2_{(1-\alpha,k-r-1)}$

Teste de Aderência - Intervalo de chegada



Intervalo	Fo	Fe	D
0 ⊣ 1	10	10	0
1 ⊣ 2	11	10	0.1
2 ⊣ 3	11	10	0.1
3 ⊣ 4	9	10	0.1
4 ⊣ 5	11	10	0.1
5 ⊣ 6	9	10	0.1
6 ⊣ 7	9	10	0.1
7 ⊣ 8	9	10	0.1
8 ⊣ 9	11	10	0.1
9 ⊣ 10	10	10	0
			0.8

Teste de Aderência - Intervalo de Chegada

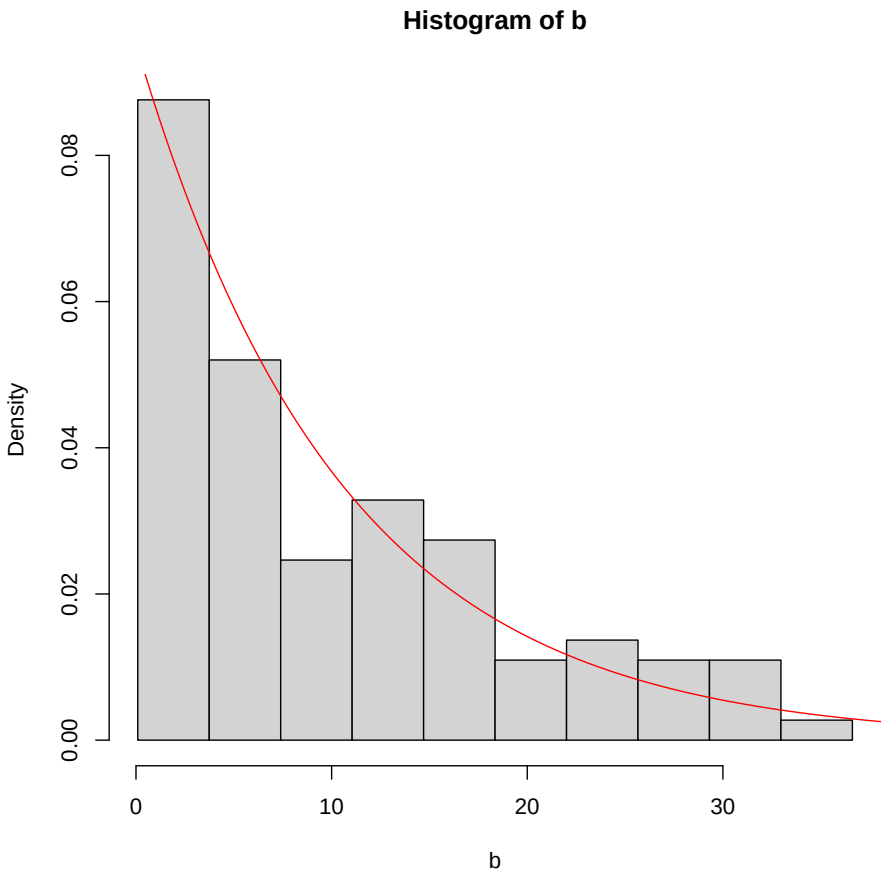
Ao nível de significância de $\alpha = 0.10$,

- Hipótese. a amostra segue a distribuição uniforme.
- Para a distribuição uniforme não há parâmetros a serem estimados a partir da amostra, então $r = 0$.
- Da tabela: $\chi^2_{0.9,9} = 14.684$.
- O valor calculado foi $D = 0.8$
- Como $D < 14.684$, a hipótese não pode ser rejeitada.

Tabela de Distribuição de Chi-Quadrado:

n \p	0.005	0.010	0.050	0.100	0.200	0.500	0.800	0.900	0.950	0.990	0.995
10	2.156	2.558	3.940	4.865	6.179	9.342	13.442	15.987	18.307	23.209	25.188
9	1.735	2.088	3.325	4.168	5.380	8.343	12.242	**14.684**	16.919	21.666	23.589
8	1.344	1.647	2.733	3.490	4.594	7.344	11.030	13.362	15.507	20.090	21.955
7	0.989	1.239	2.167	2.833	3.822	6.346	9.803	12.017	14.067	18.475	20.278
6	0.676	0.872	1.635	2.204	3.070	5.348	8.558	10.645	12.592	16.812	18.548
5	0.412	0.554	1.145	1.610	2.343	4.351	7.289	9.236	11.070	15.086	16.750
4	0.207	0.297	0.711	1.064	1.649	3.357	5.989	7.779	9.488	13.277	14.860
3	0.072	0.115	0.352	0.584	1.005	2.366	4.642	6.251	7.815	11.345	12.838
2	0.010	0.020	0.103	0.211	0.446	1.386	3.219	4.605	5.991	9.210	10.597
1	0.000	0.000	0.004	0.016	0.064	0.455	1.642	2.706	3.841	6.635	7.879

Teste de Aderência - Bytes transmitidos



Intervalo	Fo	Fe	D
0.090 ⊣ 3.743	32	29.15	0.28
3.743 ⊣ 7.396	19	20.58	0.12
7.396 ⊣ 11.049	9	14.53	2.10
11.049 ⊣ 14.702	12	10.25	0.30
14.702 ⊣ 18.355	10	7.24	1.05
18.355 ⊣ 22.008	4	5.11	0.24
22.008 ⊣ 25.661	5	3.61	0.54
25.661 ⊣ 29.314	4		↓
29.314 ⊣ 32.967	4		↓
32.967 ⊣ 36.620	1		↓
25.661 ⊣ 36.6204	9	5.61	2.05
			6.68

Teste de Aderência - Bytes transmitidos

Ao nível de significância de $\alpha = 0.10$,

- Hipótese. a amostra segue a distribuição exponencial.
- Para a distribuição exponencial há um estimador usado (média), então $r = 1$.
- Além disso, optei por agrupar as três últimas classes porque o número de ocorrencias é pequeno (Fo e Fe). O número de classes usadas foi $k = 8$.
- Da tabela: $\chi^2_{0.9,6} = 12.017$.
- O valor caldulado foi $D = 6.68$
- Como $D < 10.645$, a hipótese não pode ser rejeitada.

Tabela de Distribuição de Chi-Quadrado:

n \p	0.005	0.010	0.050	0.100	0.200	0.500	0.800	0.900	0.950	0.990	0.995
10	2.156	2.558	3.940	4.865	6.179	9.342	13.442	15.987	18.307	23.209	25.188
9	1.735	2.088	3.325	4.168	5.380	8.343	12.242	14.684	16.919	21.666	23.589
8	1.344	1.647	2.733	3.490	4.594	7.344	11.030	13.362	15.507	20.090	21.955
7	0.989	1.239	2.167	2.833	3.822	6.346	9.803	12.017	14.067	18.475	20.278
6	0.676	0.872	1.635	2.204	3.070	5.348	8.558	**10.645**	12.592	16.812	18.548
5	0.412	0.554	1.145	1.610	2.343	4.351	7.289	9.236	11.070	15.086	16.750
4	0.207	0.297	0.711	1.064	1.649	3.357	5.989	7.779	9.488	13.277	14.860
3	0.072	0.115	0.352	0.584	1.005	2.366	4.642	6.251	7.815	11.345	12.838
2	0.010	0.020	0.103	0.211	0.446	1.386	3.219	4.605	5.991	9.210	10.597
1	0.000	0.000	0.004	0.016	0.064	0.455	1.642	2.706	3.841	6.635	7.879

Exemplo

No exemplo anterior, você consegue determinar:

- Qual a capacidade mínima do enlace para este servidor?
- Quais fatores podem afetar este dimensionamento (para pior)?

QQ-Plot?

- O QQ-Plot (Quantile-Quantile Plot) é uma ferramenta gráfica utilizada para avaliar se dois conjuntos de dados (ou um conjunto de dados e uma distribuição teórica) vêm da mesma distribuição de probabilidade.
- É amplamente utilizado para verificar a suposição de normalidade em análises estatísticas, mas pode ser usado para comparar qualquer tipo de distribuição.
- Por que usar o QQ-Plot?
  - ✔ Permite uma comparação visual direta entre as distribuições.
  - ✔ Ajuda a identificar desvios da distribuição esperada (e.g., caudas pesadas, assimetria).
  - ✔ Complementa testes de hipótese formais (como Chi-Quadrado ou Kolmogorov-Smirnov) fornecendo insights visuais.

QQ-PLOT - A Essência: Comparando Quantis

- A ideia principal é comparar os **quantis** de duas distribuições.
- **Quantil**: Em estatística, um quantil é um valor que divide um conjunto de dados ordenados em partes iguais.
- Se duas distribuições são semelhantes, seus quantis correspondentes devem ser aproximadamente iguais.
- Existem diferentes tipos de quantis, dependendo do número de divisões realizadas na amostra:
  - ✔ Mediana: Divide os dados em duas partes iguais (o quantil de 50%).
  - ✔ Quartis: Dividem os dados em quatro partes. O primeiro quartil (Q1) é o quantil de 25%, a mediana é o segundo quartil (Q2), e o terceiro quartil (Q3) é o quantil de 75%.
  - ✔ Decis: Dividem os dados em dez partes.
  - ✔ Percentis: Dividem os dados em 100 partes. Por exemplo, o 90º percentil é o valor abaixo do qual 90% dos dados estão.
- Tipos de QQ-Plots:
  - ✔ **Dados vs. Dados:** Compara os quantis de duas amostras empíricas.
  - ✔ **Dados vs. Teórica:** Compara os quantis de uma amostra empírica com os quantis de uma distribuição teórica

QQ-PLOT: Exemplo

- Considere a quantidade de bytes transmitidos por requisição em um servidor Web. No entanto, os valores serão ordenados:

Decil										
D1(10%):	0.09	0.22	0.23	0.35	0.38	0.72	0.96	1.02	1.34	1.37
D2(20%):	1.48	1.48	1.5	1.51	1.55	1.65	1.71	2.13	2.26	2.33
D3(30%):	2.57	2.64	2.89	3	3.35	3.41	3.47	3.48	3.54	3.61
D4(40%):	3.63	3.64	4.05	4.05	4.86	4.91	5.1	5.23	5.38	5.54
D5(50%):	5.9	5.95	6.02	6.29	6.35	6.52	6.62	6.72	7.01	7.36
D6(60%):	7.38	8.3	8.62	9.15	9.62	9.63	9.86	9.97	10.2	10.6
D7(70%):	12.2	12.41	12.53	12.55	13.31	13.33	13.4	13.51	13.91	13.94
D8(80%):	14.59	14.62	15.15	15.92	15.95	15.97	16.21	16.42	16.52	16.78
D9(90%):	17.35	17.98	19.11	19.66	20.57	21.14	22.01	22.15	22.83	24.9
D10(100%):	25.12	25.94	26.71	26.79	27.39	30.34	31.01	31.06	31.31	36.62

- Calculando os limites para os deciles utilizando a média:

D1: Média entre 1.37 e 1.48	1.425
D2: Média entre 2.33 e 2.57	2.450
D3: Média entre 3.61 e 3.63	3.620
D4: Média entre 5.10 e 5.23	5.165
D5: Média entre 6.72 e 7.01	6.865
D6: Média entre 9.15 e 9.62	9.385
D7: Média entre 12.55 e 13.31	12.930
D8: Média entre 15.95 e 15.97	15.960
D9: Média entre 19.11 e 19.66	19.385
D10: Média entre 26.71 e 26.79	26.750

- Estes valores seriam plotados no eixo Y do QQ-Plot. No eixo X, o ponto correspondente será o da distribuição teórica sendo comparada.

Passo 2: Calcular as Probabilidades dos Quantis

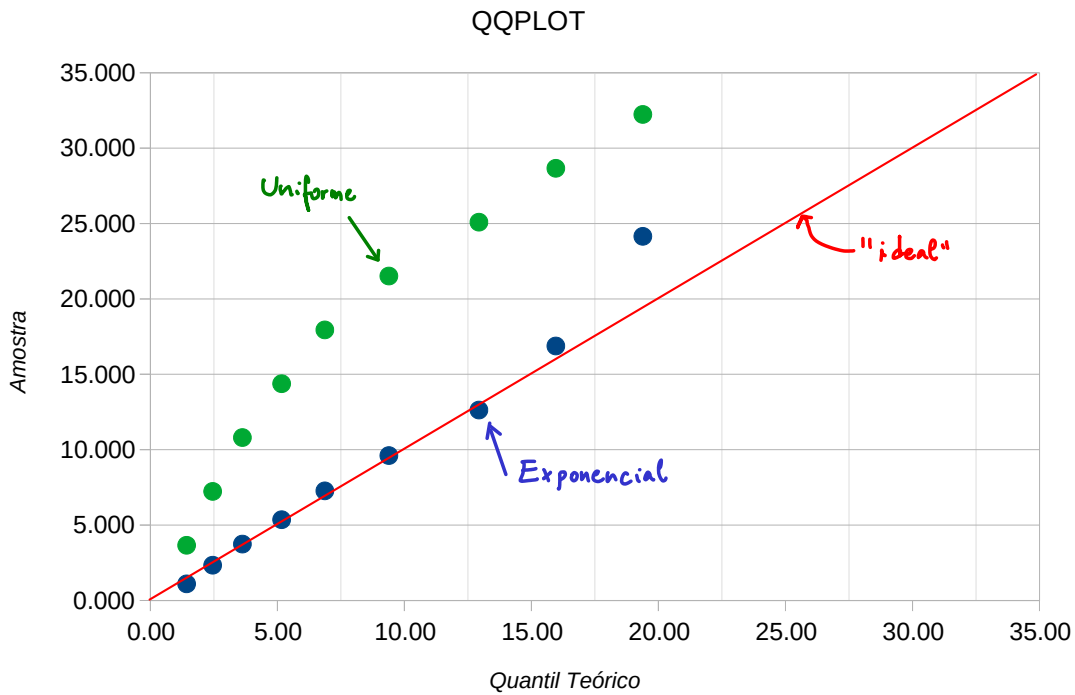
- Neste caso, a hipótese é a aderência dos dados à distribuição exponencial[2]:
  - ✔ Considere a distribuição exponencial acumulada dada por $P[X \leq x] = 1 - e^{-1/\mu x}$
  - ✔ Considere o quantil como $p = P[X \leq x]$.
  - ✔ Isolando na equação, encontramos $x = \mu \ln(1 - p)$.
- Com isso, obtemos:

Quantil Teórico (Distr. Uniforme)	Quantil Teórico (Distr. Exponencial)	Amostra (Y)
3.662	1.106	1.105
7.234	2.327	2.341
10.806	3.791	3.742
14.378	5.334	5.359
17.950	7.249	7.271
21.522	9.608	9.612
25.094	12.593	12.630
28.666	16.837	16.883
32.238	24.080	24.154
—	—	26.750

- Estes valores seriam plotados no eixo Y do QQ-Plot. No eixo X, o ponto correspondente será o da distribuição teórica sendo comparada.

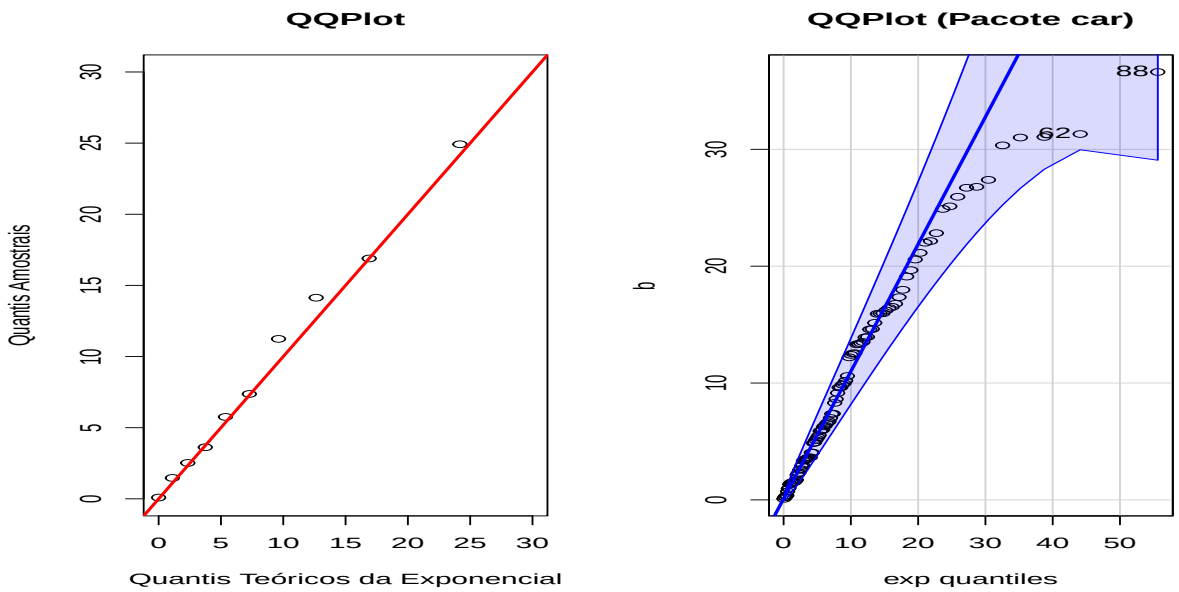
[2] adicionei a distribuição uniforme para apresentar uma comparação

Resultado: QQPlot



- Interpretação:
  - ✔ O eixo X representa os quantis teórico.
  - ✔ O eixo Y representa os dados observados.
  - ✔ Uma linha de referência $y = x$ é frequentemente adicionada ao gráfico.
  - ✔ Se os dados seguem a distribuição teórica, os pontos no QQ-Plot deveriam se alinhar aproximadamente ao longo desta linha.

Comandos R



- Comandos com o R:

```
b<-scan("dados.dat")
bq<-quantile(b,probs=seq(0,1,0.1))
quantis_teoricos <- qexp(seq(0,1,0.1), rate = 1/mean(b))
par(mfrow = c(1, 2))
qqplot(x = quantis_teoricos, y = bq,
       main = "QQPlot",
       xlab = "Quantis Teóricos da Exponencial",
       ylab = "Quantis Amostrais",
       xlim=c(0,30), ylim=c(0,30))
abline(0, 1, col = "red", lwd = 2)
library(car)
qqPlot(b,distribution="exp",rate=1/mean(b), main = "QQPlot (Pacote car)")
```

Exercício

506	487	474	352	352	577	554	341	273	669
435	555	542	249	452	485	586	422	528	502
408	551	452	468	387	447	324	543	479	489
571	617	584	530	448	533	569	724	370	433
461	471	441	392	690	384	566	630	365	654
591	438	612	423	488	558	601	444	628	525
469	531	505	537	357	486	612	529	370	451
449	346	473	453	507	456	431	429	550	410
630	582	458	517	565	531	673	541	502	639
512	448	498	587	435	583	489	373	342	470

- Realize os testes de aderência de Chi-Quadrado e QQPlot para a amostra acima

Referências


R. Jain. The art of computer systems performance analysis: techniques for experimental design, measurement, simulation and modeling. John Wiley & Sons, 1991.


V. Ricci, Fitting Distributions with R, disponível em `http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf`. 2005.