# ESTAT

## Revisões e conceitos de Estatística Descritiva

LEEC

2019/2020

## Definição: Estatística

É uma ciência dotada de um conjunto de métodos para observar, recolher, analisar e interpretar dados de fenómenos imprevisíveis tendo como objetivo auxiliar a formulação de decisões face à incerteza.

## Definição: Estatística descritiva

É um conjunto de métodos cujo objetivo consiste em resumir e representar de forma a tornar compreensível a informação contida nos dados experimentais (extração de informação).

## Definição: Estatística inferencial

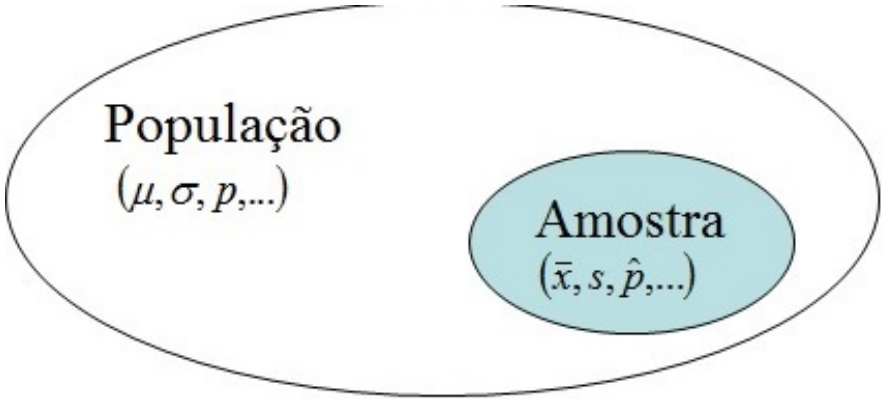
É um conjunto de métodos com o objetivo de realizar estimativas e tirar conclusões sobre uma população a partir da informação contida num subconjunto dessa população (amostra).

## Definição: População

É o conjunto de todos os objetos cuja(s) característica(s) se pretende estudar e analisar

## Definição: Amostra

É um subconjunto finito da população




Exemplo:

- População: Todos os alunos do ISEP
- Amostra: Todos os estudantes desta turma

Relativamente ao tamanho as populações podem ser:

- Finitas (número de filhos, Alunos do ISEP, ...)
- Infinitas (número de lançamentos de uma moeda até obter uma cara, tempo de espera por um autocarro)

## Definição: Censo

É o estudo de todos os elementos de uma população (só possível em populações finitas).

## Definição: Sondagem

É um estudo (da população) efetuado a partir da informação recolhida de uma amostra.

## Definição: Variável estatística

É uma variável que representa uma característica da população e passível de tomar diversos valores. Pode ser qualitativa ou quantitativa.

**Tipos de variáveis (dão origem a tipos de dados estatísticos)**

- variáveis qualitativas (dados qualitativos)
  São variáveis que traduzem uma característica não numérica
  - variáveis nominais (dados nominais, categorias)
    exemplo:sexo, cor dos olhos, raça,...
  - variáveis ordinais (dados ordinais).
    Exemplo: escolaridade (básica, secundária, superior), qualidade (má, média, boa)
- Variáveis quantitativas (dados quantitativos)
  São variáveis que resultam de processos de medição ou contagem
  - Variáveis discretas (dados discretos)
    Podem tomar um conjunto numerável ou infinitamente numerável de valores.
  - variáveis contínuas (dados contínuos)
    Podem tomar quaisquer valores num dado intervalo (podem assumir um conjunto não numerável de valores)

## Definição: Classe de uma variável estatística

É cada um dos diferentes valores que a variável pode tomar (dados quantitativos discretos), qualquer intervalo de valores (dados contínuos) ou categoria (dados qualitativos).

## Definição: Frequência absoluta da i-ésima classe ($n_i$)

É o número de observações que pertencem à classe $i$

## Definição: Frequência relativa da i-ésima classe ($f_i$)

É a quantidade

$$f_i = \frac{n_i}{n}$$

onde, $n$ o número total de observações e $c$ é o número de classes.
Verifica-se que:

$$\sum_{i=1}^{c} n_i = n, \quad \sum_{i=1}^{c} f_i = 1$$

## Definição: Frequência acumulada até à classe $i$ ($N_i$)

É o número de observações de valor inferior ou igual ao valor que caracteriza a classe $i$.

$$N_i = \sum_{j=1}^{i} n_j.$$

## Definição: Frequência relativa acumulada até à classe $i$ ($F_i$)

É a percentagem de observações de valor inferior ou igual ao valor que caracteriza a classe $i$.

$$F_i = \sum_{j=1}^{i} f_j.$$

Os dados quantitativos discretos são obtidos de observações de variáveis quantitativas discretas. Estas estão geralmente associadas a processos de **contagem**.

## Definição: Tabela de distribuição de frequências de dados quantitativos discretos

É uma tabela com $k$ linhas (uma para cada valor distinto) e três colunas. A primeira coluna representa cada uma dos diferentes valores observados, a segunda a frequência absoluta e a terceira a frequência relativa.

Exemplo: Observação de 50 peças de artesanato quanto ao número de defeitos

| Número de defeitos ($x_i$) | frequência absoluta ($n_i$) | frequência relativa ($f_i$) |
|---|---|---|
| 0 | 23 | 0.46 |
| 1 | 16 | 0.32 |
| 2 | 7 | 0.14 |
| 3 | 4 | 0.08 |

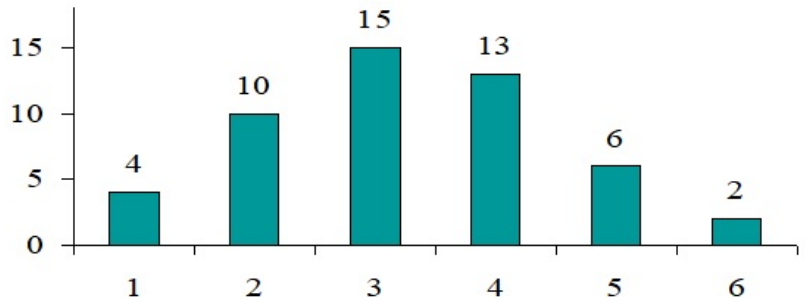
**Representação gráfica**: Gráfico de barras.

Exemplo: Dados de um inquérito realizado em 50 habitações quanto ao número de elementos do agregado familiar.

## Tabela de frequências

| Tamanho do agregado ($x_i$) | Frequência absoluta ($n_i$) | Frequência relativa ($f_i$) | Freq. Relat. acumulada ($F_i$) |
|---|---|---|---|
| 1 | 4 | 0,08 | 0,08 |
| 2 | 10 | 0,2 | 0,28 |
| 3 | 15 | 0,3 | 0,58 |
| 4 | 13 | 0,26 | 0,84 |
| 5 | 6 | 0,12 | 0,96 |
| 6 | 2 | 0,04 | 1 |
| Total | 50 | 1 | |

## Gráfico de barras



- Os dados quantitativos contínuos são obtidos a partir da observação de instâncias de uma variável numérica contínua de uma população.
- Em geral, os dados de uma variável contínua apresentam uma diversidade tal que é necessário agrupa-los em classes (intervalos).

## Definição: Classe de uma variável quantitativa contínua

É um intervalo na forma $[a, b[$ ou $]a, b]$ que representa um conjunto de valores que a variável pode tomar.

- **Amplitude da classe** $h$: $h = b - a$
- **Marca da classe** é o representante dessa classe para efeitos de cálculo. Por defeito considera-se $marca = x_i = \frac{a+b}{2}$

## Número de classes

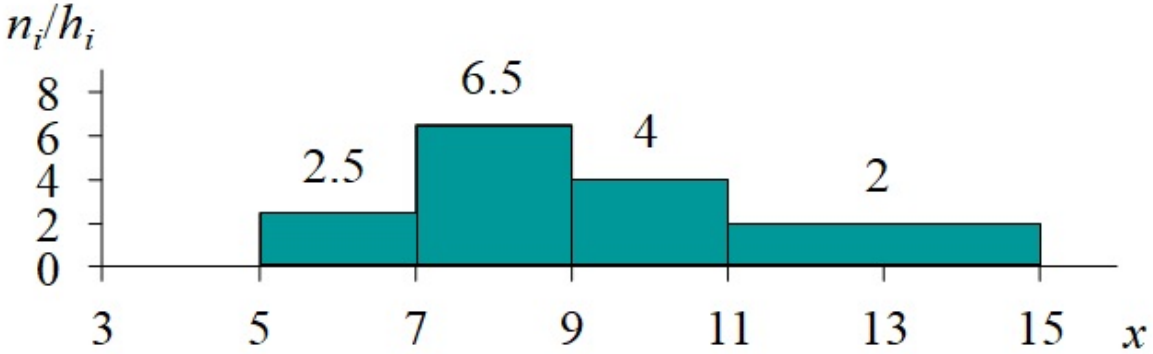
É habitual escolher entre 4 e 20 classes, dependendo do número de observações. Regra de Sturges: $c = int(1 + 3.3log(n))$.

## Representação gráfica

### Definição: Histograma

É uma representação gráfica dos dados em que se marcam as classes no eixo dos $xx$, as frequências no eixo dos $yy$ e em que se usam barras de área proporcional à frequência da classe correspondente. As barras contíguas têm fronteira comum.

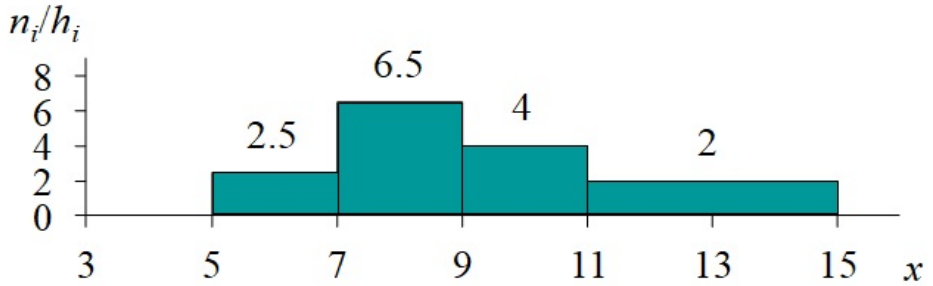
Na figura seguinte representa-se um histograma referente aotempo de realização de 34 tarefas, em minutos. Note-se que $h_i$ e $n_i$ representam, respetivamente, a amplitude e a frequência absoluta da classe $i$ : $i = 1, 2, 3, 4$.



## Tabela de frequências

| amplitude classe ($h_i$) | Tempo classe $i$ | Marca da classe($x_i'$) | Freq. abs. ($n_i$) | Freq. rel. ($f_i$) | Freq. rel. acumulada ($F_i$) |
|---|---|---|---|---|---|
| 2 | [5, 7[ | 6 | 5 | 0.1471 | 0.1471 |
| 2 | [7, 9[ | 8 | 13 | 0.3824 | 0.5295 |
| 2 | [9, 11[ | 10 | 8 | 0.2353 | 0.7648 |
| 4 | [11, 15] | 13 | 8 | 0.2353 | 1.000 |
| | | Total | 34 | 1.000 | |

## Histograma



# Algumas medidas descritivas

- Medidas de localização
  - Média
  - Mediana, Quartis, Percentis
  - Moda
- Medidas de dispersão
  - Amplitude total, amplitude inter-quartil
  - Variância e desvio padrão
  - Coeficiente de variação

## Definição: Média aritmética $\bar{x}$ para dados não classificados

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

## Definição: Média aritmética $\bar{x}$ para dados classificados

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{c} x_i n_i = \sum_{i=1}^{c} x_i f_i$$

onde $c$ é o número de classes, $n$ o número total de observações, $n_i$ é a frequência absoluta e $f_i$ a frequência relativa. No caso de dados contínuos classificados $x_i$ representa a marca da classe.

## Definição: Moda, *Mo*

É a classe (ou classes) com maior frequência. Para dados discretos, ou contínuos não classificados, é o valor (ou valores) que apresenta(m) a maior frequência.

## Definição: Mediana, *Me*

É o valor que divide uma série de $n$ observações em duas partes iguais, tal que 50% das observações tem um valor inferior, ou igual, a *Me*

- Mediana para dados não classificados em série ordenada

$$Me = \begin{cases} \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & , n \text{ par} \\ x_{(n+1)/2} & , n \text{ impar} \end{cases}$$

- Mediana para dados quantitativos contínuos classificados

$$Me = L_k + \left( \frac{0.5 - F_{k-1}}{f_k} \right) h_k$$

onde, a classe mediana é a primeira classe cuja frequência relativa acumulada ultrapassa, ou iguala 50%, $f_k$ é a frequência relativa da classe mediana, $F_{k-1}$ é a frequência relativa da classe anterior á classe mediana, $L_k$ representa o limite inferior da classe mediana e $h_k$ é a amplitude da classe mediana.

## Definição: Quantil de ordem $\alpha$, $z_\alpha$

É o valor que divide uma série de $n$ observações em duas partes, tal que $\alpha$% das observações tem um valor menor, ou igual, a $z_\alpha$

- Quantil para dados não classificados em série ordenada

$$z_\alpha = x_k$$

onde $x_k$ é o maior inteiro menor que $n\alpha + 1$

- Quantil para dados quantitativos contínuos classificados

$$z_\alpha = L_k + \left( \frac{\alpha - F_{k-1}}{f_k} \right) h_k$$

onde, $k$ é a classe do quantil $\alpha$, tal que $F_{k-1} < \alpha$ e $F_l \geq \alpha$

## Definição: Percentil de ordem $k$, $p_k$

$$p_k, (k = 1, 2, ..., 99) = z_{k/100}$$

## Definição: Decil de ordem $k$, $d_k$

$$d_k, (k = 1, 2, ..., 9) = z_{k/10}$$

## Definição: Quartil de ordem $k$, $d_k$

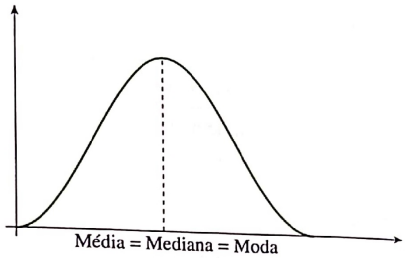
$$q_k, (k = 1, 2, 3) = z_{k/4}$$

Tendo em mente as definições realizadas anteriormente, temos que

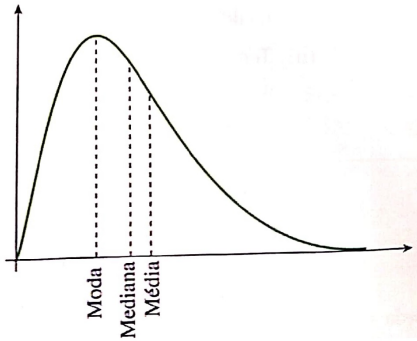
$$Me = p_{50} = d_5 = q_2$$

# Comparação entre a média moda e mediana (simetria)

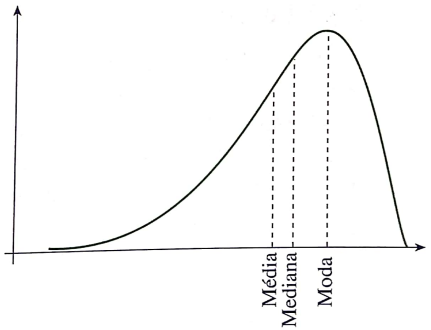
Em distribuições simétricas unimodais, média=mediana=moda.




Distribuição assimétrica positiva (enviesada à direita): moda<mediana<média.




Distribuição assimétrica negativa (enviesada à esquerda): moda>mediana>média.



## Definição: Amplitude total, "Range" $r$

É a diferença entre o maior e o menor valor de um conjunto de observações.

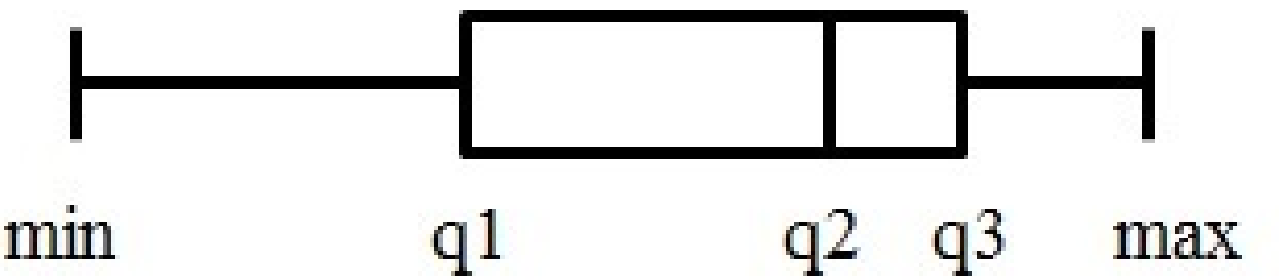
- Depende apenas das observações extremas.
- Depende o número de observações
- Para dados classificados, define-se como a diferença entre o valores máximo e mínimo das marcas da classe.

## Definição: Amplitude interquartil, "Interquartile Range" $r_q$

$$r_q = q_3 - q_1$$

## Diagrama de extremos e quartis (Boxplot ou caixa de bigodes)

Consiste numa representação gráfica de dados baseado nos seus quartis, mínimo e máximo muito utilizada para analisar a dispersão dos dados.




- O limite inferior da caixa representa o primeiro quartil
- A linha divisória da caixa a mediana e o limite superior o terceiro quartil
- A largura (altura) da caixa representa a amplitude inter-quartil (medida de dispersão)

## Outliers moderados

Um valor observado $x_i, i = 1, 2, ..., n$, é um candidato a outlier moderado se

$$x_i < q_1 - 1.5(q_3 - q_1)$$

ou

$$x_i > q_3 + 1.5(q_3 - q_1)$$

## Outliers severos

Um valor observado $x_i, i = 1, 2, ..., n$, é um candidato a outlier severo se

$$x_i < q_1 - 3(q_3 - q_1)$$

ou

$$x_i > q_3 + 3(q_3 - q_1)$$

- A remoção de outliers(valores anómalos) do conjunto de dados requer o conhecimento da área onde o estudo estatístico se insere.

A variância e o desvio padrão são as medidas de variabilidade, ou dispersão mais utilizadas em estatística.

- Estas medidas têm em conta todos os valores observados.
- O desvio padrão indica a proximidade com que os valores observados se distribuem em torno da média.
- Um valor nulo do desvio padrão implica que todas as observações concentradas em torno do mesmo valor.
- Valores crescentes do desvio padrão indicam que os valores estão cada vez mais "espalhados"ou dispersos em relação à média.
- A variância é o quadrado do desvio padrão.

## Definição: Variância de uma amostra $s^2$

Para dados não classificados: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Para dados classificados:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{c}(x_i - \bar{x})^2 n_i$$

Nota: Para dados contínuos classificados $x_i$ representa a marca da classe.

## Definição: Desvio padrão $s$

É a raiz quadrada positiva da variância

$$s = \sqrt{s^2}$$

## Definição: Coeficiente de variação $s$

O coeficiente de variação de uma amostra é dado pela expressão

$$cv = \frac{s}{\bar{x}}$$

Trata-se de uma medida de dispersão relativa que tem em conta a magnitude dos valores observados.

As medidas de localização e as medidas de dispersão embora forneçam informação importante são insuficientes para uma boa caracterização da distribuição dos dados em frequência.
É ainda necessária

- Informação sobre a deformação dos dados
- Informação sobre o peso dos dados nas caudas

Esta informação é fornecida pelas medidas de forma.
Começamos por definir momento centrado.

## Definição: Momento centrado de ordem $r$, $m_r$

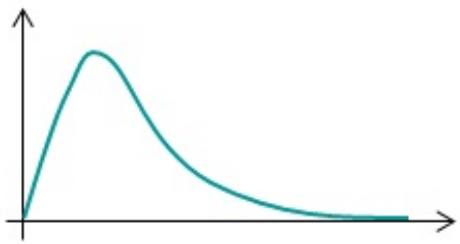
Para dados não classificados: $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$ Para dados classificados:

$$m_r = \frac{1}{n}\sum_{i=1}^{c}(x_i - \bar{x})^r n_i$$

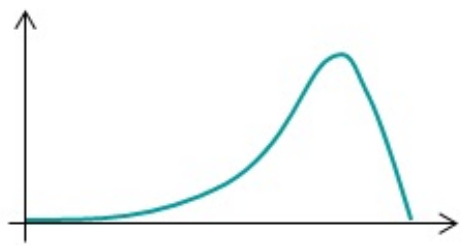
## Definição: Coeficiente de assimetria amostral, $a_3$

$$a_3 = \frac{m_3}{s^3}$$

- É uma medida adimensional
- Mede a assimetria da distribuição
- $a_3 > 0$ quando a cauda direita é mais comprida (enviesada à direita)
- $a_3 < 0$ quando a cauda esquerda é mais comprida (enviesada à esquerda)
- $a_3 = 0$ distribuição simétrica



Distribuição enviesada à direita; assimétrica positiva

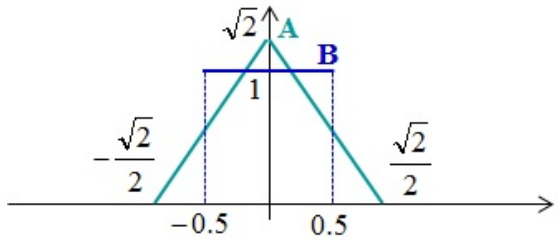


Distribuição enviesada à esquerda; assimétrica negativa

## Definição: Coeficiente de curtose amostral, $a_4$

$$a_4 = \frac{m_4}{s^4}$$

- É uma medida adimensional
- Mede o achatamento e o peso das caudas da distribuição.
- A distribuição normal tem $a_4 = 3$
- $a_4 > 3$ quando a distribuição é mais esguia e as caudas mais pesada do que a distribuição normal.
- $a_4 < 3$ quando a distribuição é mais achatada e as caudas menos pesada do que a distribuição normal.




Verifica-se que $a_4(A) > a_4(B)$ porque a distribuição B é mais achatada e tem caudas menos pesadas

Exercício 1: Dados discretos

Considere a série estatística ordenada que representa as respostas quanto so número de elementos do agregado familiar de amostra aleatória de 50 questionários:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 |
| 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 6 | 6 |

a) Construa a tabela de frequências.

b) Represente a distribuição dos dados num gráfico adequado

c) Calcule o número médio a mediana e a moda do número observado de elementos do agregado familiar.

d) Calcule a variância e o desvio padrão da amostra.

e) Classifique os dados da amostra quanto à simetria.

Exercício 2: Dados contínuos
Considere a série estatística ordenada que representa o tempo de realização de 34 tarefas, em minutos.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5.11 | 5.21 | 5.62 | 6.77 | 6.80 | 7.01 | 7.11 | 7.12 | 7.21 | 7.22 |
| 7.25 | 7.25 | 7.51 | 7.52 | 7.65 | 7.3 | 8.5 | 8.57 | 9.11 | 9.21 |
| 10.1 | 10.23 | 10.31 | 10.46 | 10.83 | 10.91 | 11.03 | 11.99 | 12.3 | 13.4 |
| 14.5 | 14.69 | 14.89 | 14,91 | | | | | | |

Considerando os dados classificados nas classes [5,7[,[7,9[,[9,11[,[11,15]:
a) Construa a tabela de frequências.
b) Represente a distribuição dos dados num histograma.
c) Calcule o tempo médio de realização das tarefas com base na tabela de frequências.
d) Represente a função cumulativa e deduza por leitura gráfica o valor da mediana.
e) Identifique a classe modal e deduza por leitura gráfica uma aproximação pontual.
f) Calcule a variância e o desvio padrão da amostra.
g) Classifique os dados da amostra quanto à simetria e curtose.

# Bibliografia

- Pedrosa, A. e Gama, S. (2004). Introdução Computacional à Probabilidade e Estatística. Porto Editora. ISBN: 972-0-06056-5

- Montegomery and Runger, Applied Statistics and Probability for Engineers, 4th Ed, John Wiley and Sons, 2007