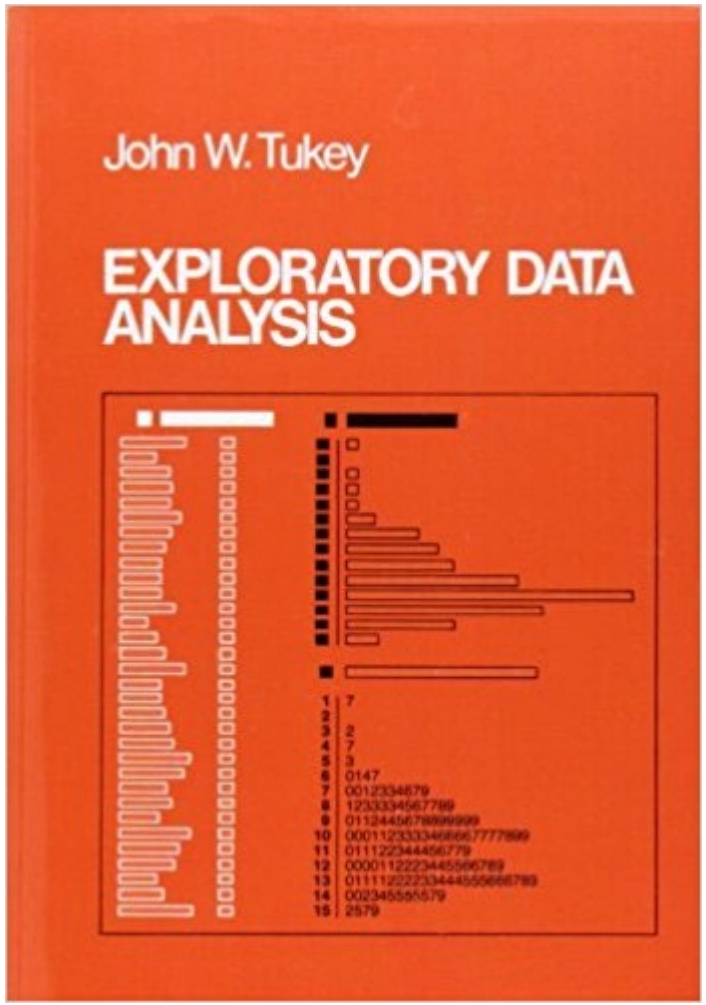
Técnicas Estatísticas de Caracterização e Visualização de Dados

PME3463 Introdução à Qualidade
Escola Politécnica da Universidade de São Paulo
Departamento de Engenharia Mecânica
Prof. Dr. Walter Ponge-Ferreira

Algumas técnicas

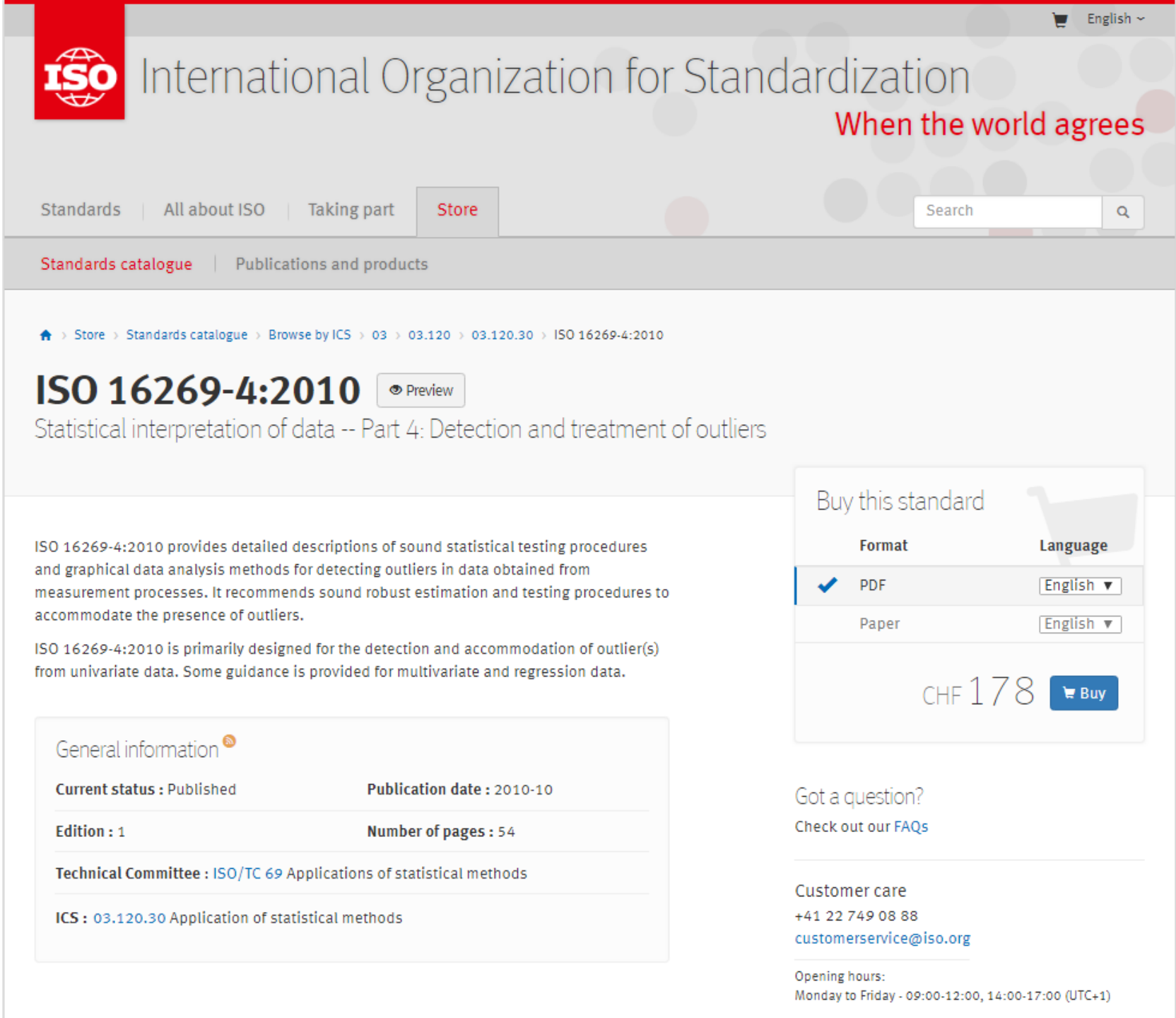
- Diagrama de pontos
- Diagrama de Ramos-e-Folhas
- Diagrama de Dobradiças
- Diagrama de Cinco Números
- Diagrama de Caixa-e-Bigodes
- Histograma
- Teste de normalidade
- Teste de outliers

Exploratory Data Analysis
by John W. Tukey



ISO 16269-4:2010
Statistical interpretation of data -- Part 4: Detection and treatment of outliers

ISO 16269-4:2010 provides detailed descriptions of sound statistical testing procedures and graphical data analysis methods for detecting outliers in data obtained from measurement processes. It recommends sound robust estimation and testing procedures to accommodate the presence of outliers.
ISO 16269-4:2010 is primarily designed for the detection and accommodation of outlier(s) from univariate data. Some guidance is provided for multivariate and regression data.



ISO 5479:1997
Statistical interpretation of data -- Tests for departure from the normal distribution

English

International Organization for Standardization

When the world agrees

Standards | All about ISO | Taking part | Store

Search

Standards catalogue | Publications and products

Store › Standards catalogue › Browse by ICS › 03 › 03.120 › 03.120.30 › ISO 5479:1997

ISO 5479:1997

Preview

Statistical interpretation of data -- Tests for departure from the normal distribution

This standard was last reviewed and confirmed in 2006. Therefore this version remains current.

General information

Current status : Published

Publication date : 1997-05

Edition : 1

Number of pages : 33

Technical Committee : ISO/TC 69 Applications of statistical methods

ICS : 03.120.30 Application of statistical methods | 17.020 Metrology and measurement in general

Buy this standard

Format	Language
✔ PDF	English
Paper	English

CHF 138 Buy

Got a question?

Check out our FAQs

Customer care

+41 22 749 08 88
customerservice@iso.org

Opening hours:
Monday to Friday - 09:00-12:00, 14:00-17:00 (UTC+1)

2.6
resistant estimation
estimation method that provides results that change only slightly when a small portion of the data values in a **data set** (2.1) is replaced, possibly with very different data values from the original ones
2.7
robust estimation
estimation method that is insensitive to small departures from assumptions about the underlying probability model of the data
NOTE An example is an estimation method that works well for, say, a **normal distribution** (2.22), and remains reasonably good if the actual distribution is skew or heavy-tailed. Classes of such methods include the L-estimation [weighted average of **order statistics** (2.10)] and M-estimation methods (see Reference [9]).

Caracterização da População (amostra)

- Medidas de posição
  - Tendência central
- Dispersão
- Simetria
- Forma
- Elementos anômalos (outliers)

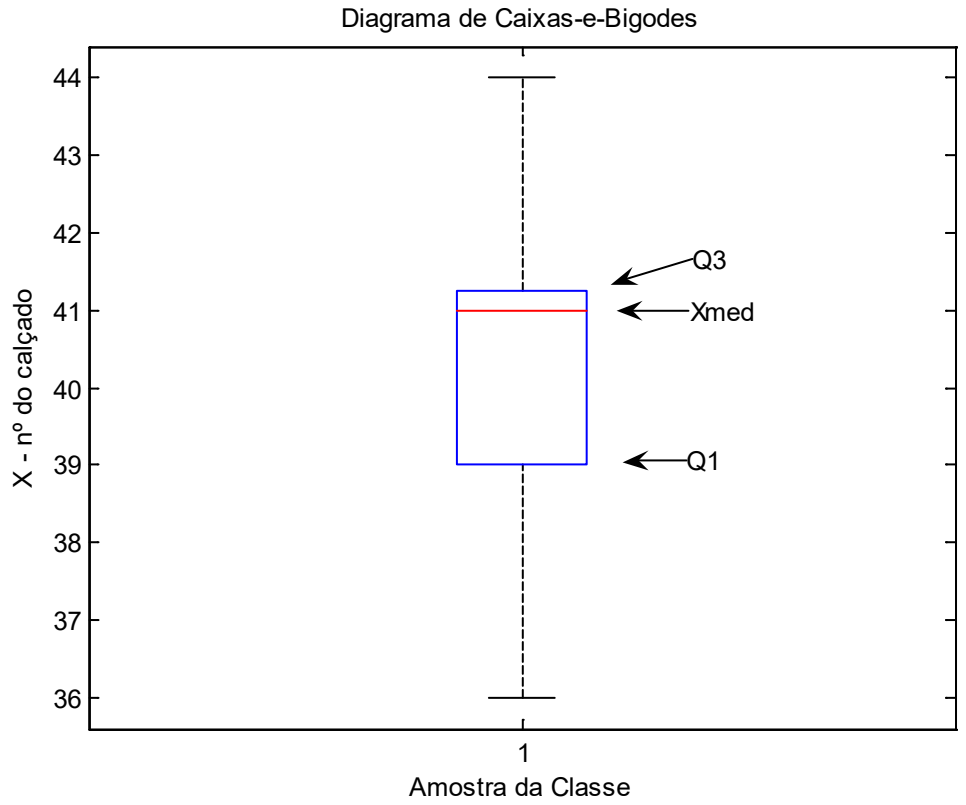
- Distribuição de probabilidade
  - Variáveis discretas: distribuição de probabilidade (Função de Repartição)
  - Variáveis contínuas: função de densidade de probabilidade

Diagrama de Caixa-e-Bigodes

- Mediana: Xmed = Q_2
- 1º e 3º Quartis: Q_1 e Q_3
- Distância interquartis: DIQ = $Q_3 - Q_1$
- Comprimento max. dos bigodes: $\Delta = 1{,}5$ DIQ
- Limite Superior: $L_s = Q_3 + \Delta$
- Limite Inferior: $L_i = Q_1 - \Delta$
- Bigode Superior: $\max\{x_i \mid x_i \leq L_s\}$
- Bigode Inferior: $\min\{x_i \mid x_i \geq L_i\}$

Diagrama de Caixa-e-Bigodes

i	x_i	x_j	
1	36	36	Xmin
2	40	38	
3	40	39	
4	38	39	Q1
5	42	40	
6	39	40	
7	41	41	Q2
8	41	41	
9	41	41	
10	42	41	Q3
11	44	42	
12	41	42	
13	39	44	Xmax
Xmax =	44		
xmin =	36		
R =	8		
Xmed =	41		
Q_1 =	39		
Q_3 =	41		
DIQ =	2		
Δ =	3		
Ls =	44		
Li =	36		



2.16
box plot
horizontal or vertical graphical representation of the **five-number summary** (2.15).
NOTE 1 For the horizontal version, the **first quartile** (2.12) and the **third quartile** (2.13) are plotted as the left and right sides, respectively, of a box, the **median** (2.11) is plotted as a vertical line across the box, the whiskers stretching downwards from the first quartile to the smallest value at or above the **lower fence** (2.17) and upwards from the third quartile to the largest value at or below the **upper fence** (2.18), and value(s) beyond the lower and upper fences are marked separately as **outlier(s)** (2.2). For the vertical version, the first and third quartiles are plotted as the bottom and the top, respectively, of a box, the median is plotted as a horizontal line across the box, the whiskers stretching downwards from the first quartile to the smallest value at or above the lower fence and upwards from the third quartile to the largest value at or below the upper fence and value(s) beyond the lower and upper fences are marked separately as outlier(s).

NOTE 2 The box width and whisker length of a box plot provide graphical information about the location, spread,
skewness, tail lengths, and outlier(s) of a sample. Comparisons between box plots and the density function of a) uniform,
b) bell-shaped, c) right-skewed, and d) left-skewed distributions are given in the diagrams in Figure 1. In each distribution,
a histogram is shown above the boxplot.
NOTE 3 A box plot constructed with its **lower fence** (2.17) and **upper fence** (2.18) evaluated by taking k to be a value
based on the sample size n and the knowledge of the underlying distribution of the sample data is called a modified box
plot (see example, Figure 2). The construction of a modified box plot is given in 4.4.

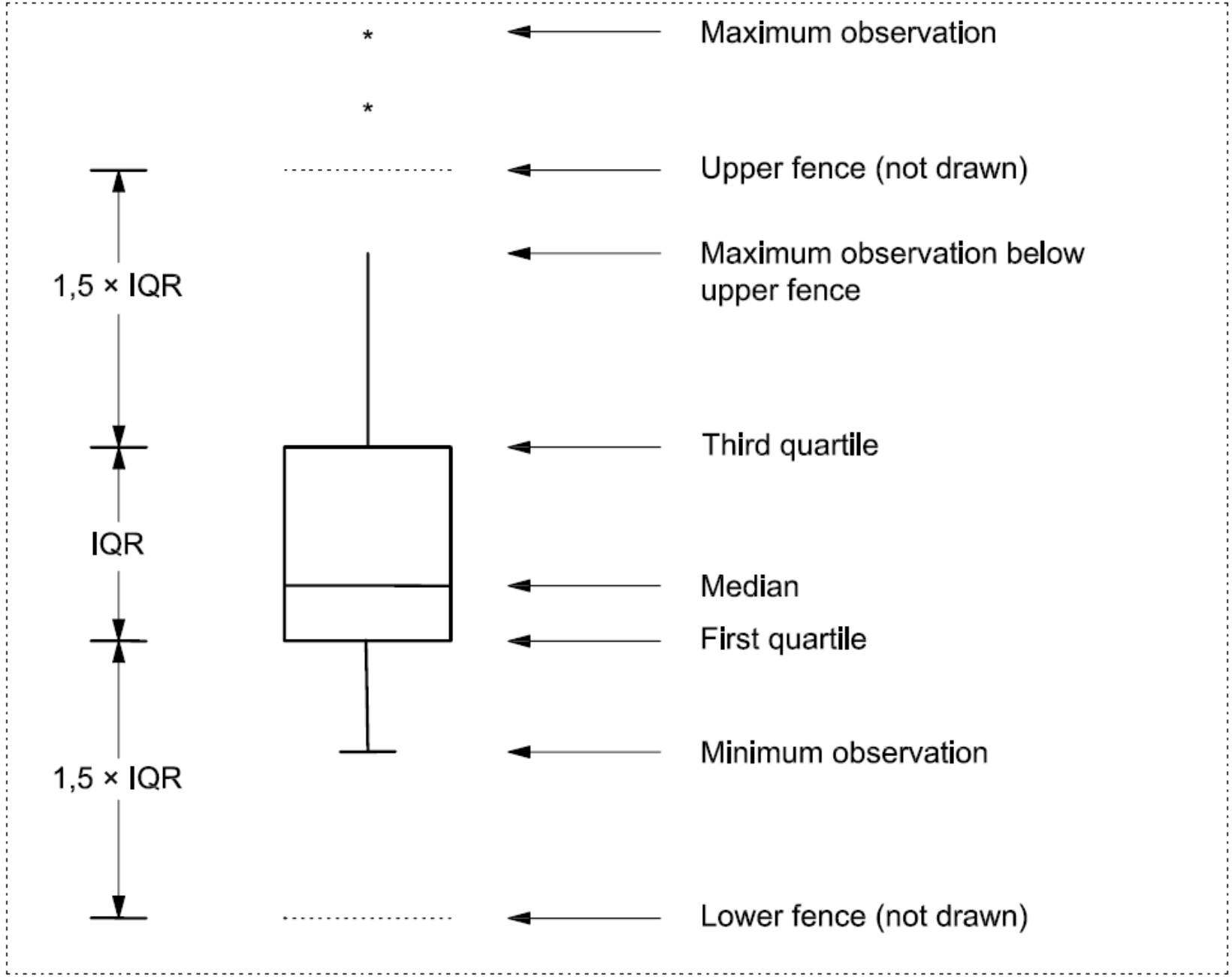


Figure 2 — Modified box plot with lower and upper fences

Diagrama de Ramos-e-Folhas

- Ordenar os elementos da amostra
- Apresentação visual da distribuição de valores
- Ordena-se os algarismos mais significativos dos elementos em colunas, chamadas de ramos.
- Ordena-se os algarismos menos significativos na horizontal, chamados de folhas.
- Exemplo: {36;38;39;40;40;41;41;41;41;42;42;44}

Histograma

- Tamanho da amostra: n
- Amplitude: $A = \max(x_i) - \min(x_i)$
- Número de faixas: $k \approx \sqrt{n}$
- Largura de faixa: $\Delta \approx \frac{A}{k}$
- Limites das Faixas:

$$\min(x_i); \min(x_i) + \Delta; \min(x_i) + 2\Delta; \ldots; \min(x_i) + (k-1)\Delta$$

- Freqüência:

$$f_i = cont(x_i) \quad | \{\min(x_i) + (i-1)\Delta \leq x_i < \min(x_i) + i\Delta\}$$

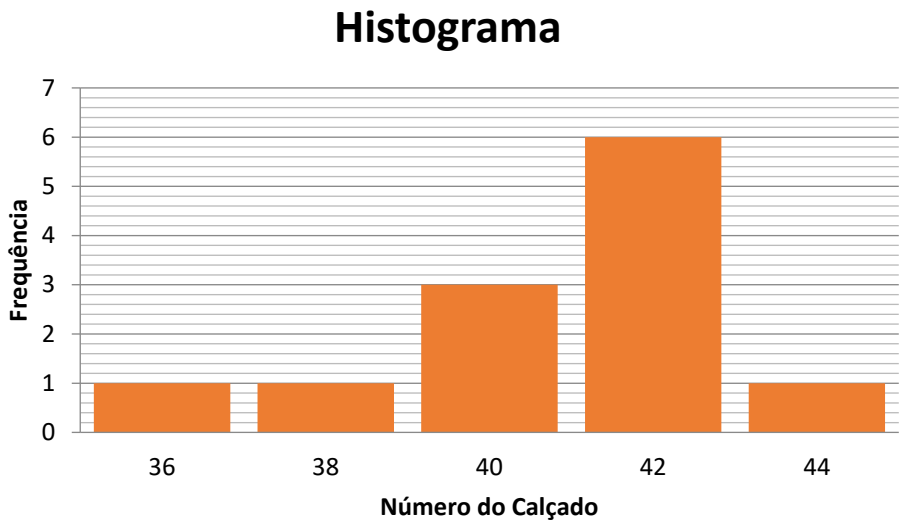
- Proporção (freqüência relativa): $p_i = \frac{f_i}{n}$

Exemplo Histograma: v.a. discreta

- Idade dos alunos da amostra:

$$n = 13 \quad \rightarrow \quad k \approx \sqrt{13} = 3{,}61 \approx 4 \quad \rightarrow \quad \Delta \approx \frac{A}{k} = \frac{44 - 36}{4} = \frac{8}{4} = 2$$

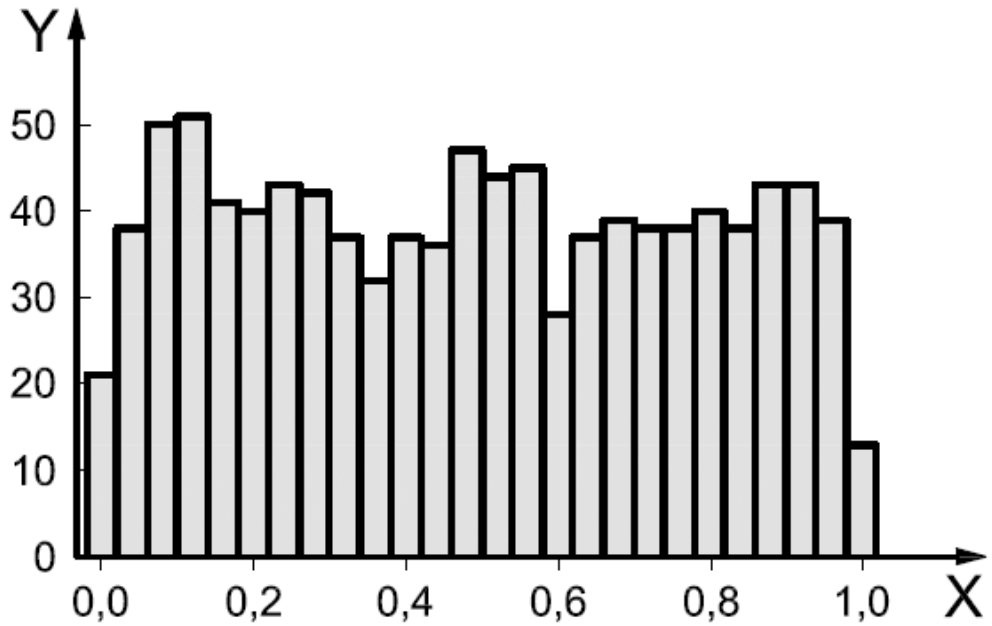
X	fi
34	0
36	1
38	1
40	3
42	6
44	1

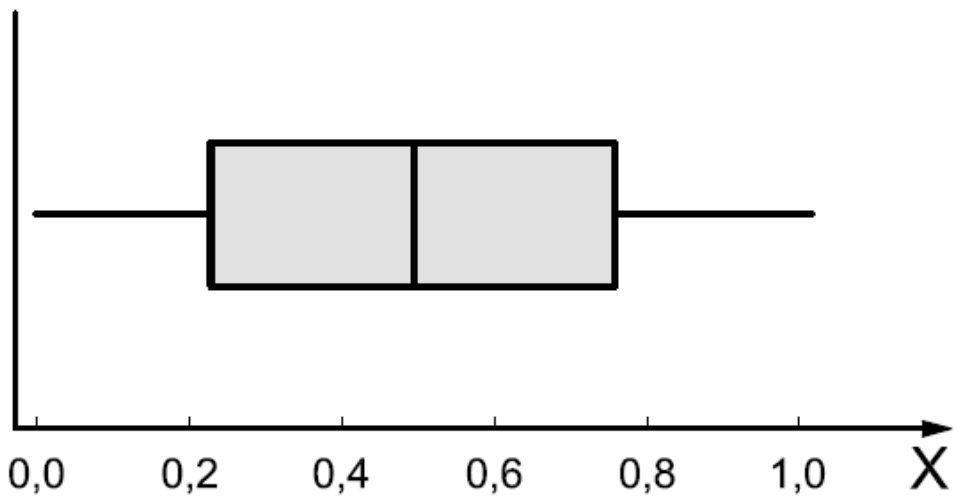


Exemplo Histograma: v.a. discreta

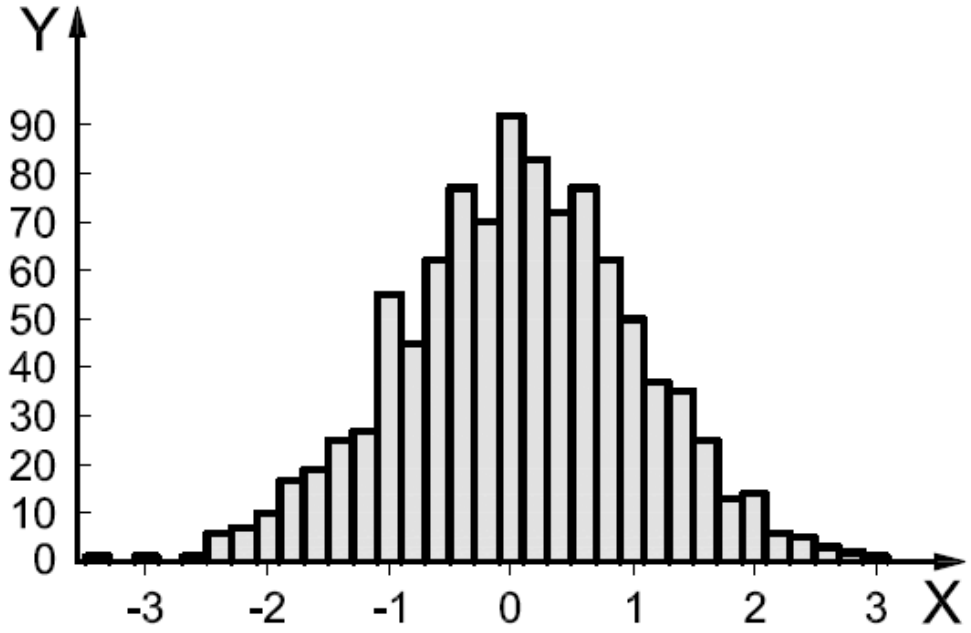
- Idade dos alunos da amostra:

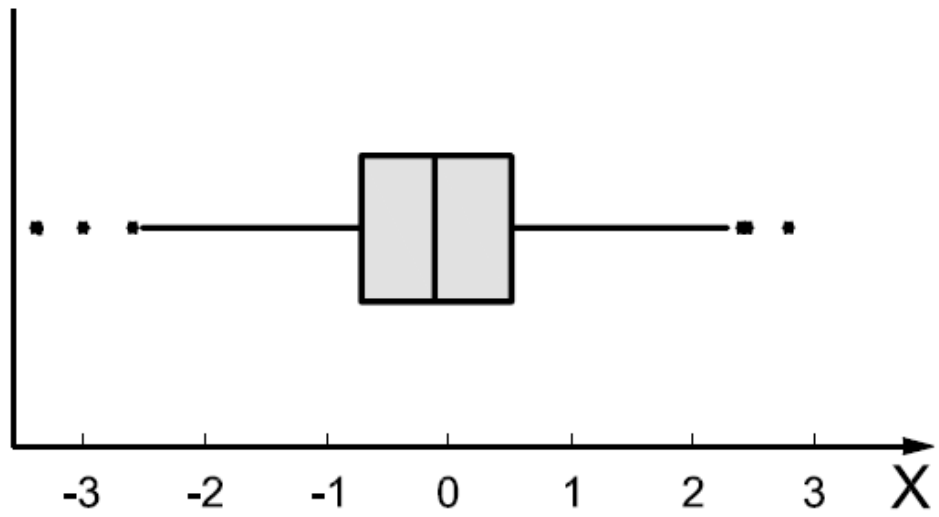
$$n = 10 \quad \rightarrow \quad k \approx \sqrt{10} = 3{,}16 \approx 4 \quad \rightarrow \quad \Delta \approx \frac{A}{k} = \frac{57 - 21}{4} = \frac{36}{4} = 9 \approx 10$$



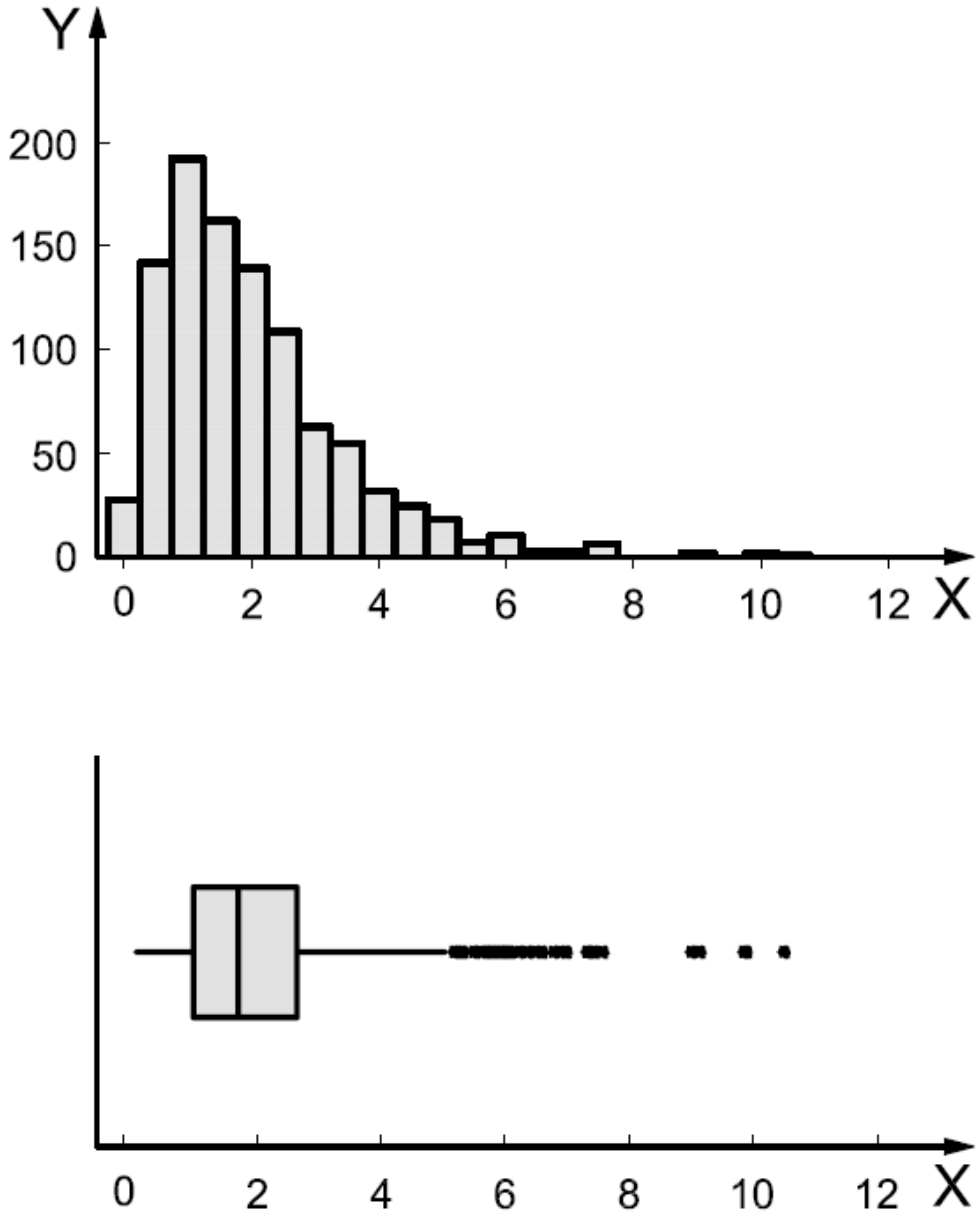


a) Uniform distribution

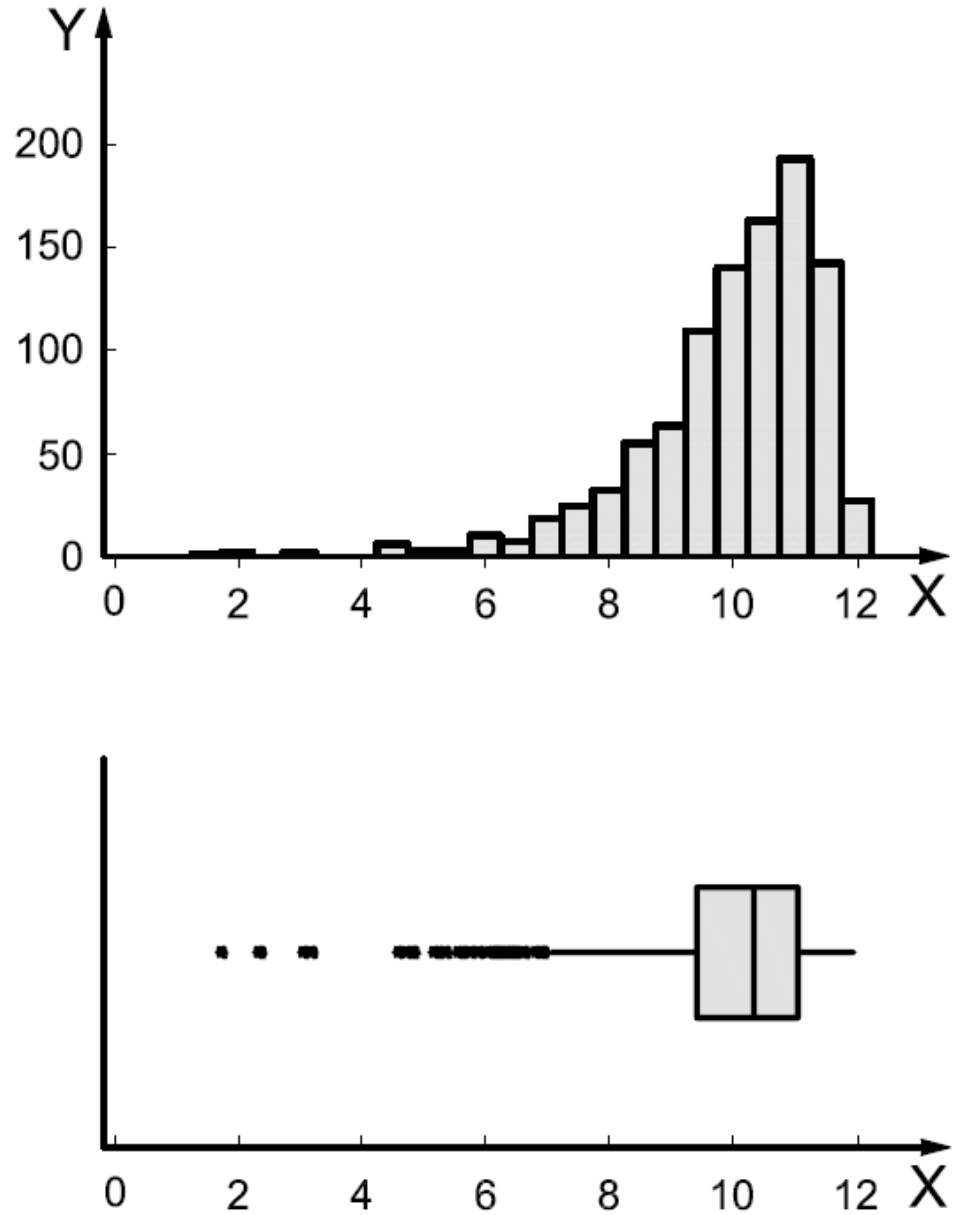




b) Bell-shaped distribution



c) **Right-skewed distribution**



d) **Left-skewed distribution**

ABNT NBR ISO 3534-1:2010
Estatística – Vocabulário e símbolos
Parte 1: Termos estatísticos gerais e termos usados em probabilidade

NORMA BRASILEIRA

ABNT NBR ISO 3534-1

Primeira edição
12.05.2010

Válida a partir de
12.06.2010

Estatística — Vocabulário e símbolos
Parte 1: Termos estatísticos gerais e termos usados em probabilidade

Statistics – Vocabulary and symbols
Part 1: General statistical terms and terms used in probability

ICS 01.040.03; 03.120.30

ISBN 978-85-07-02066-0

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS

Número de referência
ABNT NBR ISO 3534-1:2010
69 páginas

1.20
coeficiente de assimetria amostral
média aritmética da terceira potência das **variáveis aleatórias amostrais padronizadas** (1.19) de uma **amostra aleatória** (1.6)

EXEMPLO Continuando com o exemplo de 1.9, o coeficiente de assimetria amostral observado pode ser calculado como sendo 0,971 88. Para um tamanho de amostra de 10, como neste exemplo, o coeficiente de assimetria amostral é altamente variável e portanto deve ser usado com cuidado. Utilizando a fórmula alternativa da Nota 1, o valor calculado é 1,349 83.

NOTA 1 A fórmula correspondente à definição é

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{S}\right)^3$$

Alguns programas estatísticos usam a seguinte fórmula para a correção de **tendência** (1.33) do coeficiente de assimetria amostral:

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n} Z_i^3$$

onde

$$Z_i = \frac{X_i - \overline{X}}{S}$$

Para um tamanho de amostra grande, a distinção entre as duas estimativas é insignificante. A razão da estimativa não tendenciosa em relação à tendenciosa é 1,389 para n = 10, 1,031 para n = 100 e 1,003 para n = 1 000.

1.21
coeficiente de curtose amostral
média aritmética da quarta potência das **variáveis aleatórias amostrais padronizadas** (1.19) de uma **amostra aleatória** (1.6)

EXEMPLO Continuando com o exemplo de 1.9, o coeficiente de curtose amostral observado pode ser calculado como 2,674 19. Para um tamanho de amostra de 10, como neste exemplo, o coeficiente de curtose amostral é altamente variável, desta forma deve ser usado com cuidado. Programas estatísticos utilizam vários ajustes no cálculo do coeficiente de curtose amostral (ver Nota 3 de 2.40). Utilizando a fórmula alternativa indicada na Nota 1, o valor calculado é 0,436 05. Os dois valores 2,674 19 e 0,436 05 não são diretamente comparáveis. Para tanto, tomar 2,67419 - 3 (para relacionar à curtose da distribuição normal que é 3), resultando em - 0,325 81, podendo ser apropriadamente comparado a 0,436 05.

NOTA 1 A fórmula que corresponde à definição é:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{S}\right)^4$$

Para a correção do **bias** (1.33) do coeficiente de curtose amostral e para indicar o desvio da curtose da distribuição normal (que é igual a 3), alguns programas estatísticos usam a seguinte fórmula:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}Z_i^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

onde

$$Z_i = \frac{X_i - \overline{X}}{S}$$

O segundo termo na expressão é aproximadamente 3 para valores elevados de *n*. Às vezes a curtose é expressa como um valor, como definido em 2.40, menos 3 para enfatizar comparações com a distribuição normal. Obviamente, os usuários de programas estatísticos precisam estar cientes se houver ajustes nos cálculos.

2.39
coeficiente de assimetria

γ_1

momento de ordem 3 (2.34) na **distribuição de probabilidade padronizada** (2.32) de uma **variável aleatória** (2.10)

NOTA 1 Uma definição equivalente é baseada na **esperança matemática** (2.12) da terceira potência de $(X - \mu)/\sigma$, denotada por $E\left[(X - \mu)^3 / \sigma^3\right]$

NOTA 2 O **coeficiente de assimetria** é uma medida da simetria de uma **distribuição** (2.11) e é denotado às vezes por $\sqrt{\beta_1}$. Para **distribuições** simétricas, o coeficiente de assimetria é igual a 0 (supondo que os momentos apropriados na definição existam). Exemplos de distribuições com assimetria igual a zero incluem a **distribuição normal** (2.50), a **distribuição beta** (2.59) para $\alpha = \beta$ e a **distribuição de** t (2.53), contanto que os momentos existam.

2.40
coeficiente de curtose

β_2

momento de ordem 4 (2.34) na **distribuição de probabilidade padronizada** (2.32) de uma **variável aleatória** (2.10)

EXEMPLO Continuando com o exemplo da bateria de 2.1 e 2.7, para calcular o coeficiente de curtose, notar que

$$E\left\{[X - E(X)]^4\right\} = E\left(X^4\right) - 4\,E\,(X)\,E\left(X^3\right) + 6[E(X)]^2\,E\left(X^2\right) - 3\,[E(X)]^4$$

NOTA 2 O coeficiente de curtose é uma medida da densidade das caudas de uma **distribuição** (2.11). Para a **distribuição uniforme** (2.60), o coeficiente de curtose é 1,8; para a **distribuição normal** (2.50), o coeficiente de curtose é 3; para a **distribuição exponencial** (2.58), o coeficiente de curtose é 9.

NOTA 3 Deve-se ter cuidado ao se considerar valores relatados de curtose, porque alguns usuários subtraem 3 (a curtose da distribuição normal) do valor que é calculado pela definição.