

# **Introdução Estatística descritiva**

## Introdução

### Objecto da Estatística

**Recolha, compilação, análise e interpretação da informação**

*Recolha da informação* → Amostragem

*Descrição, classificação  
e apresentação da informação* → Estatística Descritiva

*Interpretação da informação* → Inferência Estatística

## Amostragem

O processo de amostragem deve ser objectivo e não tendencioso pelo que o melhor critério é o da amostragem aleatória. Dessa forma, garante-se que todos os elementos da população têm igual hipótese de ser integrados na amostra.

## Estatística Descritiva

**Síntese e representação de forma comprehensível da informação contida num conjunto de dados (*construção de tabelas, gráficos ou cálculo de medidas centrais, de dispersão ou outras*).**

## Inferência Estatística

A partir de um conjunto limitado de dados (*amostra*), pretende-se caracterizar o todo a partir dos quais os dados foram obtidos (*população*).

## População

Caracteriza-se pelo grupo inteiro de objectos (unidades) dos quais se pretende obter a informação.

## Unidade

Qualquer membro individual da população.

## Amostra

Uma parte ou subconjunto da população usada para obter informação acerca do todo.

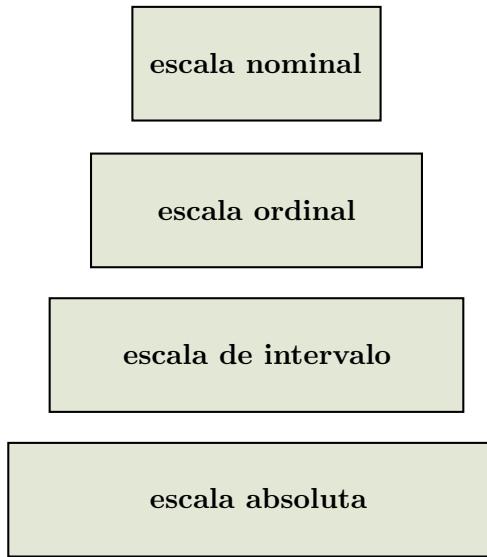
## Variável

Uma característica de uma unidade, que será medida a partir da unidade da amostra.

## Escalas de Representação

DGI

2019

**dados qualitativos**

Dados classificados por categorias não ordenadas

*ex. género, cor do cabelo, cor dos olhos*

Dados classificados por categorias ordenadas

*ex. notas num teste: mau, medíocre, suf, bom e muito bom*

Dados expressos numa escala numérica com origem arbitrária

*ex. intervalo de temperatura*

Dados expressos numa escala numérica com origem fixa

*ex. peso expresso em kg*

**dados quantitativos**

## Tipo de dados

**Qualitativos** → **Discretos**

**Quantitativos** → **Discretos**

→ **Contínuos**

## 1.1. Dados Qualitativos

### 1.1.1. Representação tabular/gráfica dos dados

A descrição das amostras faz-se com recurso a **tabelas de frequência, a diagramas de barras e a diagramas circulares.**

#### Cálculo de Frequências

$$F_k$$

frequência absoluta da categoria  $k$

$$F'_k = \sum_{k=1}^K F_k$$

frequência absoluta acumulada

$$f_k = \frac{F_k}{n}$$

frequência relativa da categoria  $k$

$$f'_k = \sum_{k=1}^K f_k = \frac{F'_k}{n}$$

frequência relativa acumulada

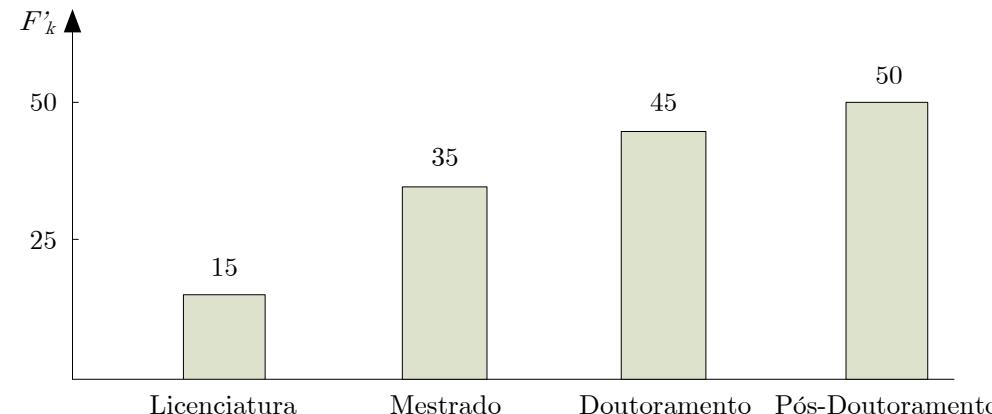
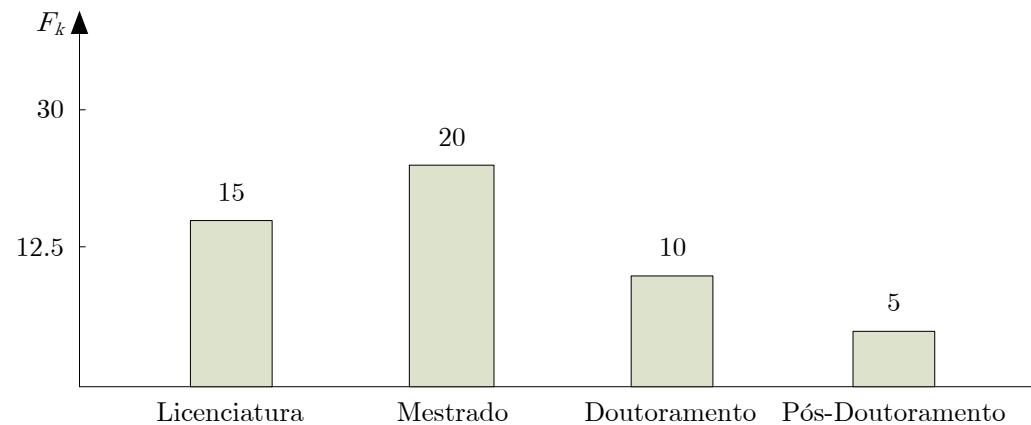
## Exemplo

Numa amostra constituída por 50 professores de uma instituição de Ensino Superior constatou-se que 15 tinham como habilitação o grau de Licenciado, 20 o grau de Mestre, 10 o grau de Doutor e 5 encontravam-se em trabalhos de Pós-Doutoramento.

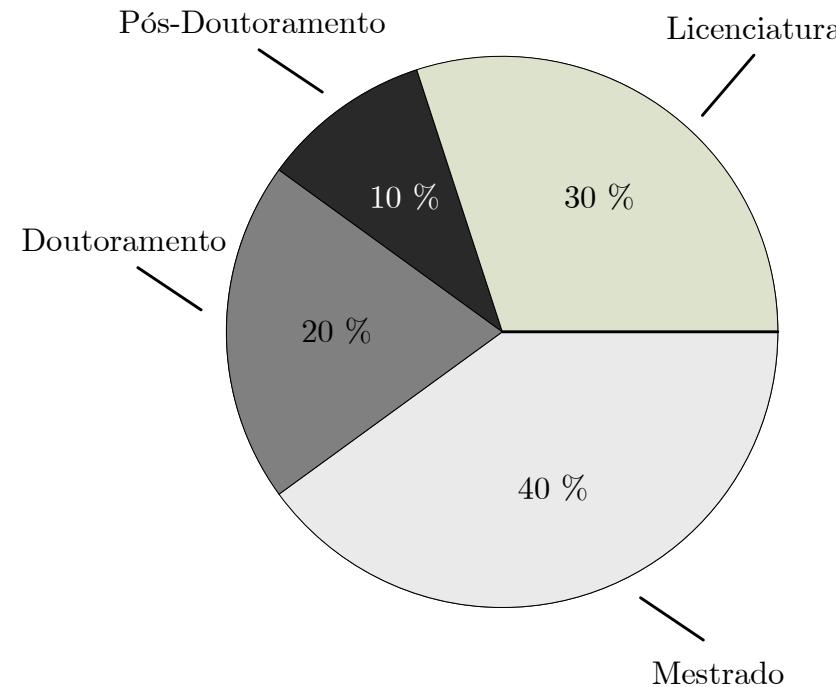
- ▶ A variável *grau académico* é uma variável *qualitativa* e pode ser representada numa *escala ordinal*.

### i) Tabela de frequências

Grau académico	$F_k$	$F'_k$	$f_k$	$f'_k$
<b>Licenciatura</b>	15	15	0.3	0.3
<b>Mestrado</b>	20	35	0.4	0.7
<b>Doutoramento</b>	10	45	0.2	0.9
<b>Pós-Doutoramento</b>	5	50	0.1	1.0

ii) *Diagramas de barras*

iii) *Diagrama circular*



## 1.2. Dados Quantitativos

### 1.2.1. Representação tabular/gráfica dos dados

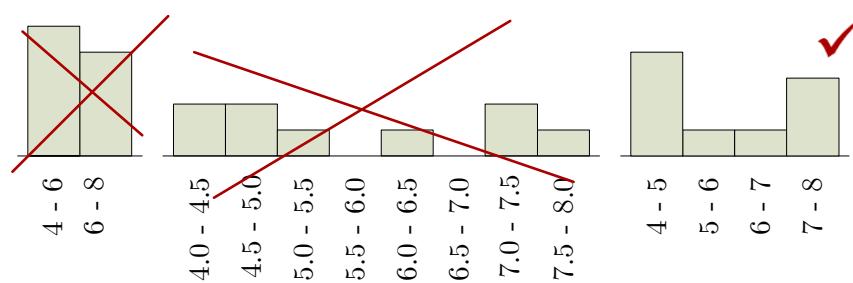
#### Construção de Histogramas

DGI

2019

Denomina-se de histograma a representação gráfica dos dados em que se marcam as classes no eixo horizontal, as frequências no eixo vertical e se usam barras de área proporcional à frequência da classe correspondente.

► *O número de intervalos ou classes considerado não deve ser demasiado pequeno para não esconder a variabilidade, nem demasiado grande, para se poder evidenciar a regularidade.*



$$\text{Regra prática - nº classes} \approx \sqrt{n}$$

(classes de igual amplitude)

## Exemplo

Considere as classificações obtidas por 49 alunos num exame de Estatística:

9.2	5.2	10.8	10.1
8.3	13.2	9.1	13.5
9.4	12.5	8.4	11.1
8.5	9	11.3	11.9
6.2	8.1	9.5	13.3
14	5.7	7.7	6.8
13.8	6.7	8.3	12.9
9.2	7.4	9.4	6.7
9.9	7.8	4.3	10
9.8	12.3	9.8	8.5
10.3	8.1	6.5	
15	7.1	11.7	
12.9	13.6	11.2	

máximo

15

mínimo

Construa a tabela de frequências, o histograma e o respectivo polígono de frequências.

Classe		$F_i$	$f_i$
[ 0, 1 [		0	0
[ 1, 2 [		0	0
[ 2, 3 [		0	0
[ 3, 4 [		0	0
[ 4, 5 [		1	0.02
[ 5, 6 [		2	0.041
[ 6, 7 [		5	0.102
[ 7, 8 [		4	0.082
[ 8, 9 [		7	0.143
[ 9, 10 [		10	0.204
[ 10, 11 [		4	0.082
[ 11, 12 [		5	0.102
[ 12, 13 [		4	0.082
[ 13, 14 [		5	0.102
[ 14, 15 [		1	0.02
[ 15, 16 [		1	0.02
[ 16, 17 [		0	0
[ 17, 18 [		0	0
[ 18, 19 [		0	0
[ 19, 20 ]		0	0
	$\sum$	49	1

### Definição das classes do histograma

$$\min\{x_1, x_2, \dots, x_{49}\} = 4.3 \quad \rightarrow \quad A = 15 - 4.3 = 10.7$$

$$\max\{x_1, x_2, \dots, x_{49}\} = 15.0$$

*Utilização de números redondos*

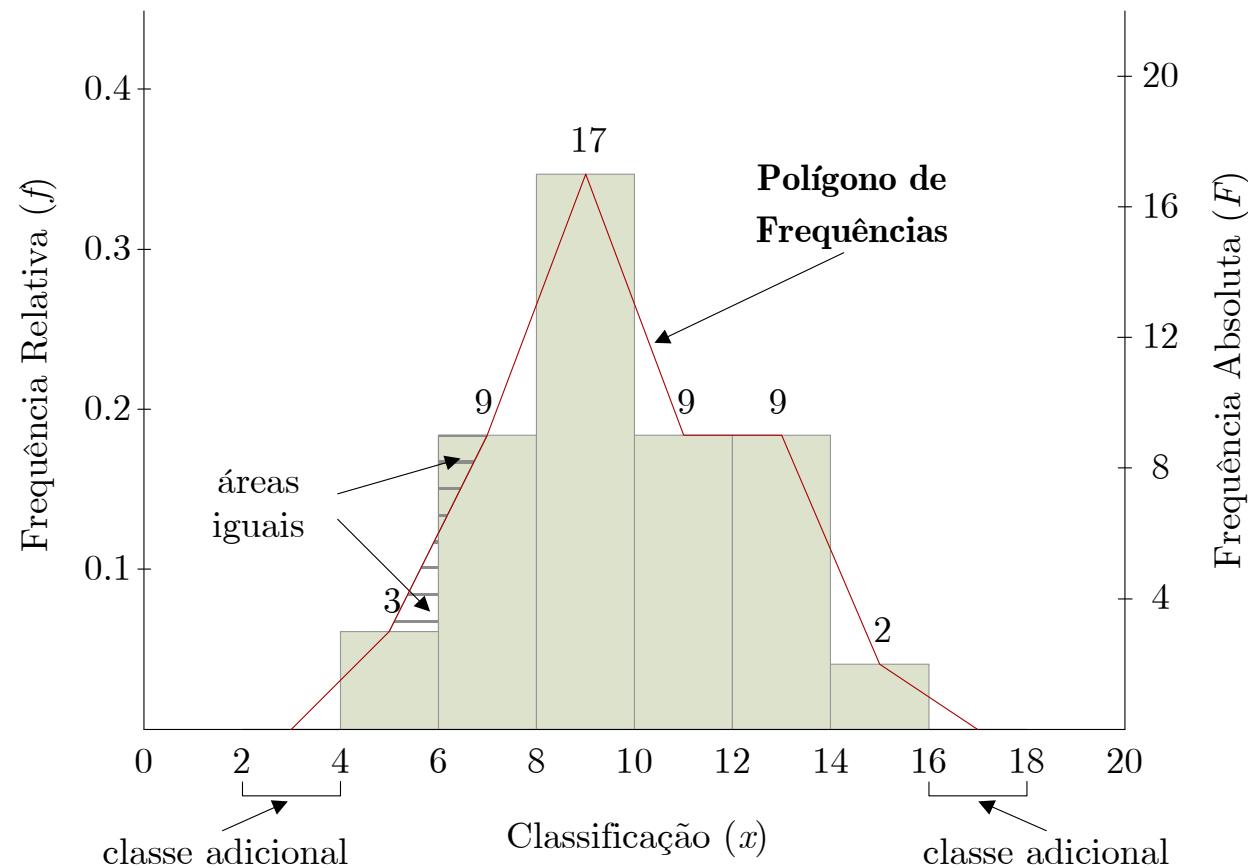
$$\begin{aligned} \text{nº classes} &\approx \sqrt{n} = \sqrt{49} = 7 & \rightarrow & \text{amplitude das classes} = 2 \\ A_k &= 10.7/7 = 1.529 & & \text{valor mínimo} = 4 \end{aligned} \quad \left. \begin{array}{l} \\ \end{array} \right\} 6 \text{ classes}$$

Classe		$F_k$	$F'_k$	$f_k$	$f'_k$
[ 4, 6 [		3	3	0.061	0.061
[ 6, 8 [		9	12	0.184	0.245
[ 8, 10 [		17	29	0.347	0.592
[ 10, 12 [		9	38	0.184	0.776
[ 12, 14 [		9	47	0.184	0.959
[ 14, 16 ]		2	49	0.041	1
$\sum$		49		1	

## Histograma e Polígono de Frequências

DGI

2019



### 1.2.2. Estatísticas de Localização

#### 1.2.2.1 Média Amostral ( $\bar{x}$ )

- A média toma um *valor* que é *central* em relação aos dados que constituem a amostra.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

- A média *minimiza a soma dos erros quadráticos* dos dados.

$$SEQ = \sum_{i=1}^n (x_i - c)^2$$

## Expressões de Cálculo da Média Amostral

DGI

2019

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

*dados não agrupados*

$$\bar{x} = \frac{1}{n} \sum_{k=1}^K F_k \times x_k \quad \text{ou} \quad \bar{x} = \sum_{k=1}^K f_k \times x_k \quad \text{dados discretos agrupados}$$

$$\bar{x} \approx \frac{1}{n} \sum_{k=1}^K F_k \times M_k \quad \text{ou} \quad \bar{x} \approx \sum_{k=1}^K f_k \times M_k \quad \text{dados contínuos agrupados}$$

## Exemplo

Calcule a classificação média obtida pelos 49 alunos no exame de Estatística.

DGI

2019

i) *Dados não agrupados*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{49} \times 476 = 9.7143$$

ii) *Dados agrupados*

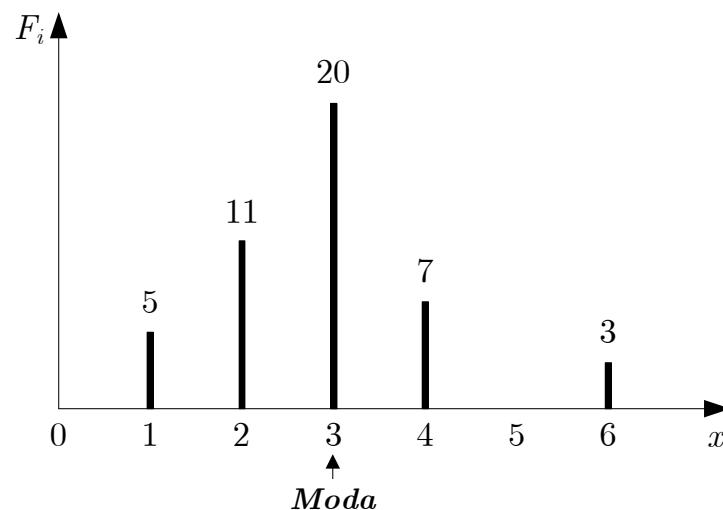
$$\begin{aligned}\bar{x} &\approx \frac{1}{n} \sum_{k=1}^K F_k \times M_k = \frac{1}{49} (3 \times 5 + 9 \times 7 + 17 \times 9 + 9 \times 11 + 9 \times 13 + 2 \times 15) = \\ &= 9.7347\end{aligned}$$

### 2.2.2.2. Moda (Mod)

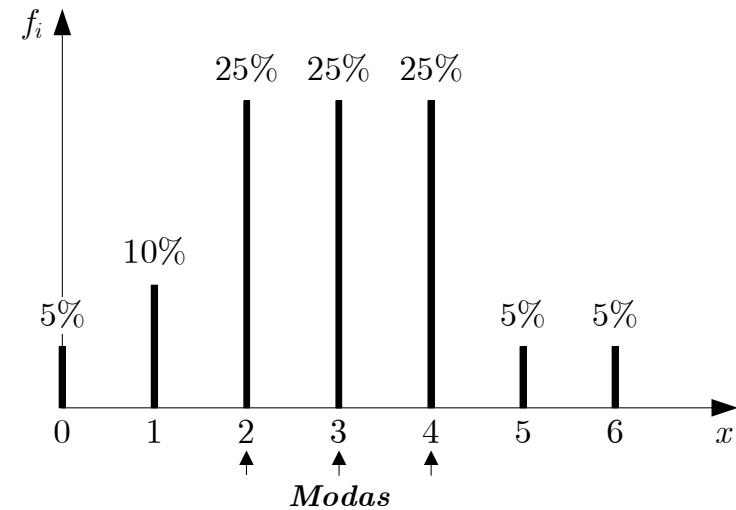
**Valor mais comum de um conjunto de observações.**

- ▶ Pode não existir (conjunto amodal) e se existir pode não ser única.

#### i) Dados Discretos



(Distribuição unimodal)

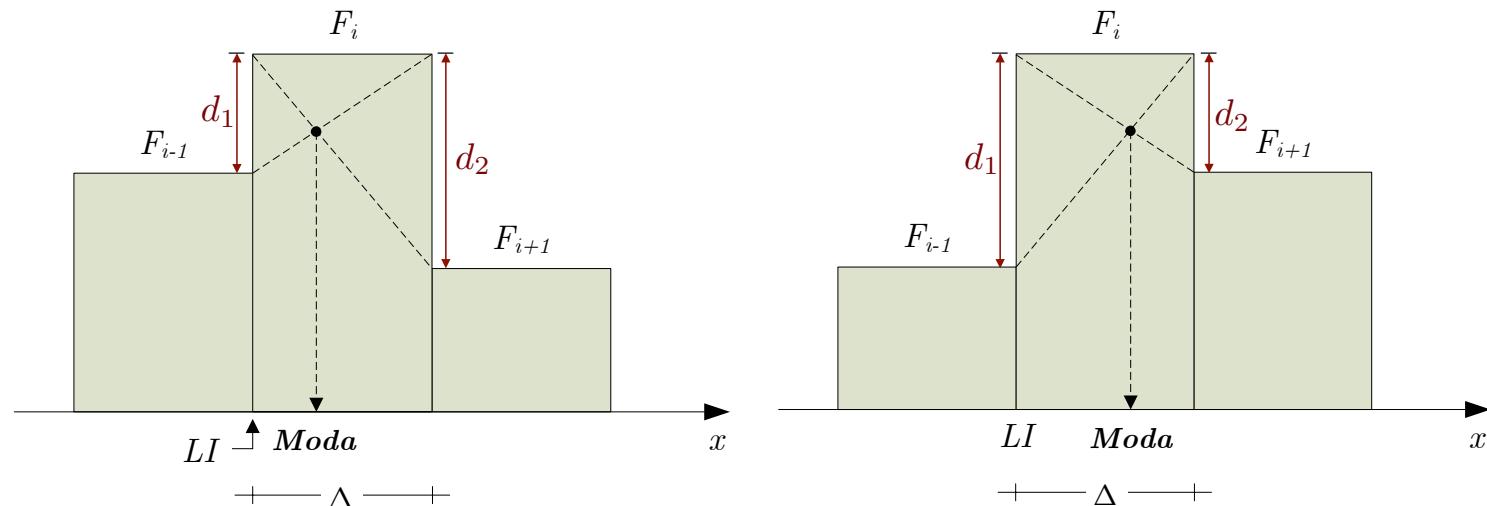


(Distribuição multimodal)

## ii) Dados Contínuos

- Identificação da classe (ou classes) a que corresponde a maior frequência - Classe modal.

$$Moda = LI + \frac{d_1}{d_1 + d_2} \times \Delta$$



(aproxima a moda à classe adjacente de maior frequência)

### 1.2.2.3. Mediana (Med)

**A ideia é “partir ao meio” o conjunto dos valores observados.**

- ▶ Inicia-se o cálculo com a *ordenação dos dados* por ordem crescente ou decrescente formando o vector:  $(x_{1:n}, x_{2:n}, \dots, x_{n:n})$

#### Dados Discretos

##### i) Dados não agrupados

- ▶ **Se  $n$  for ímpar, a mediana toma o valor do dado que, nesse vector, ocupa a posição central:**

$$Med = x_{(n+1)/2}$$

#### Exemplo

Calcule a mediana para o seguinte conjunto de valores: {6, 5, 8, 8, 4, 3, 10}

- ***Se  $n$  for par, a mediana toma o valor médio dos dois termos cuja localização no vector se aproxima mais da posição central:***

$$Med = \frac{x_{n/2} + x_{n/2+1}}{2}$$

### Exemplo

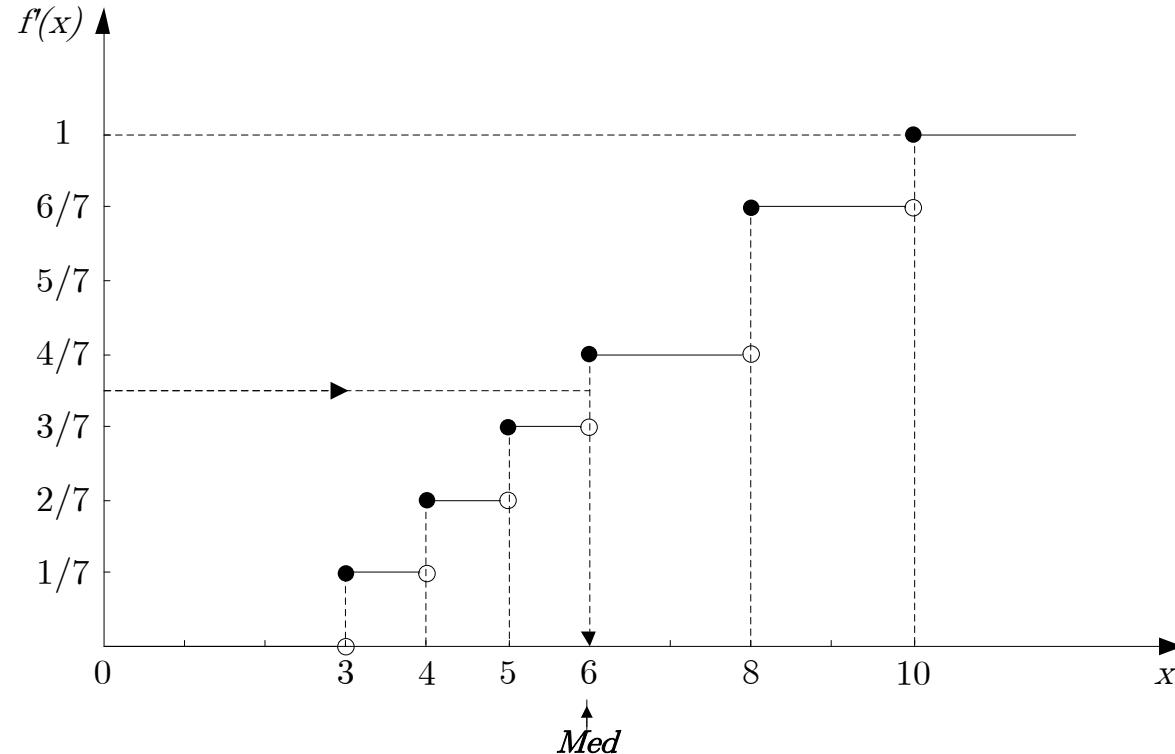
Calcule a mediana do seguinte conjunto de valores: {1, 3, 3, 5, 7, 8}

## Dados agrupados

{3, 4, 5, 6, 8, 8, 10}

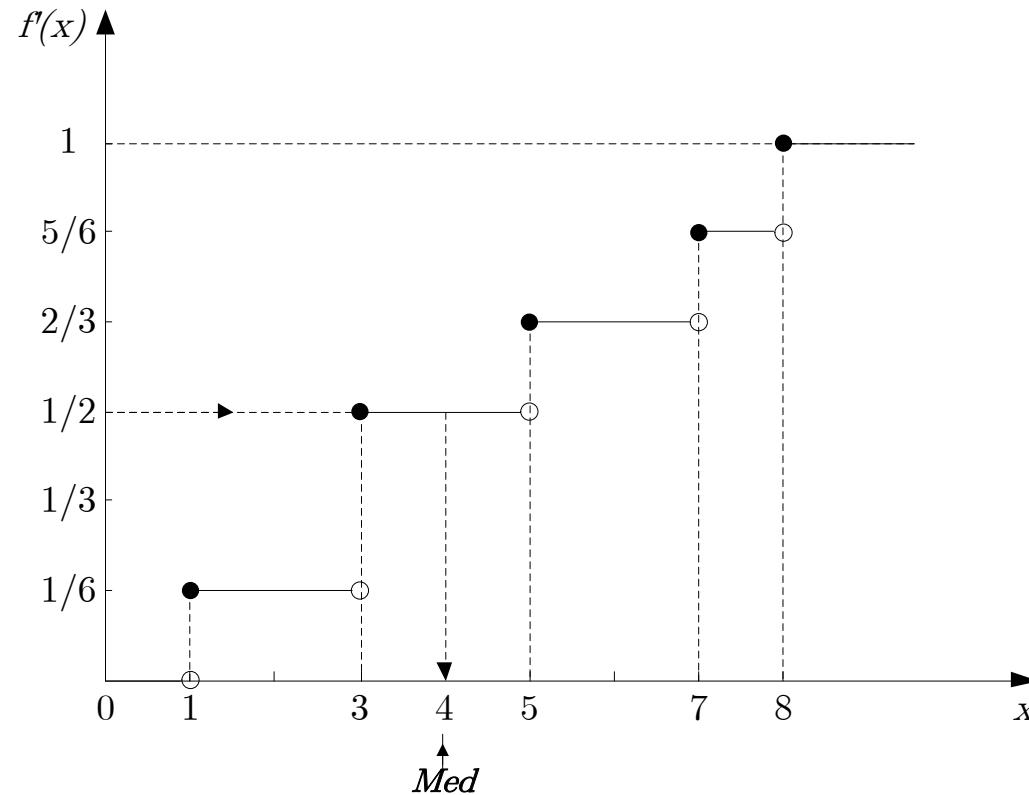
DGI

2019



## Dados agrupados

$$\{1, 3, 3, 5, 7, 8\}$$



## Dados Contínuos

- ▶ Para dados contínuos agrupados, começa por fazer-se a *identificação da classe mediana*.
- ▶ A classe mediana é aquela onde as frequências acumuladas passam de um valor inferior para um valor superior a metade dos dados ( $n/2$  ou 0.5)

$$Med = LI + \frac{n/2 - F'_{i-1}}{F_i} \times \Delta$$

(frequências absolutas)

$$Med = LI + \frac{0.5 - f'_{i-1}}{f_i} \times \Delta$$

(frequências relativas)

Sendo:

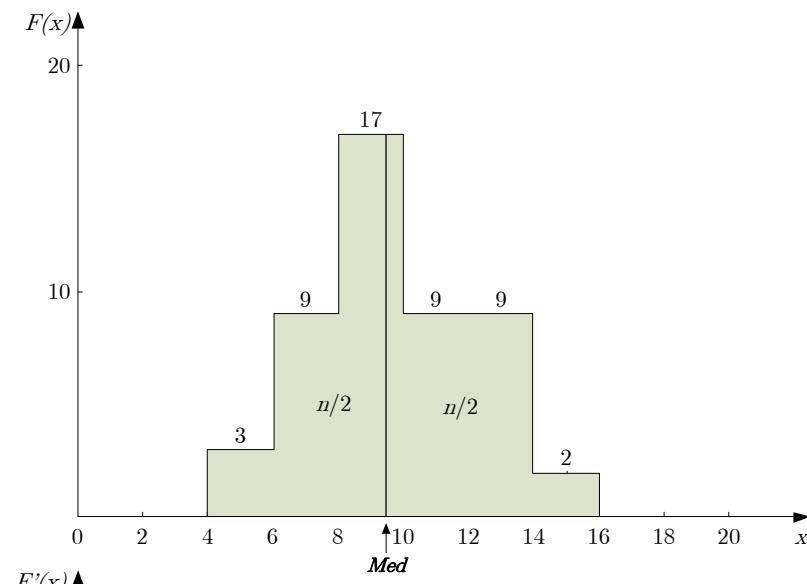
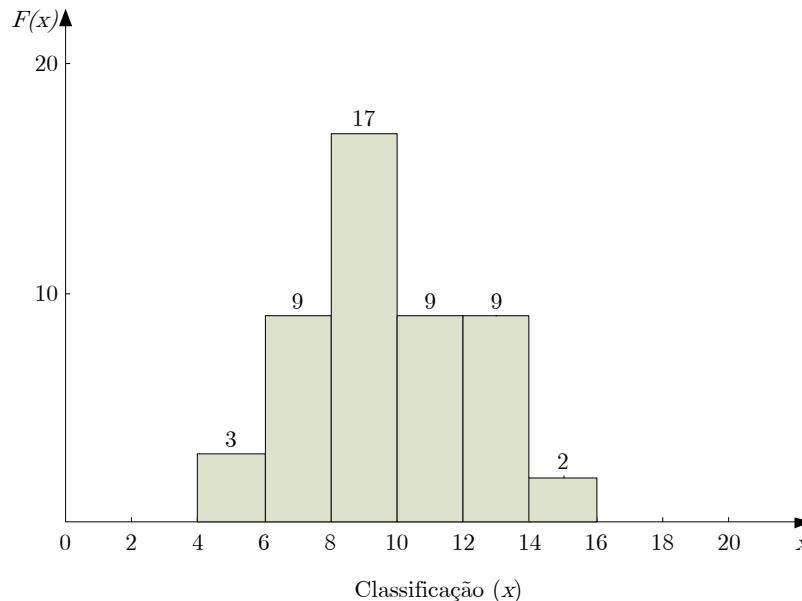
$LI$  - limite inferior da classe mediana

$F'_i$  e  $f'_i$  - frequências absoluta e relativa acumuladas da classe mediana

$F'_{i-1}$  e  $f'_{i-1}$  - frequências absoluta e relativa acumuladas da classe que precede a classe mediana

$F_i$  e  $f_i$  - frequências absoluta e relativa da classe mediana

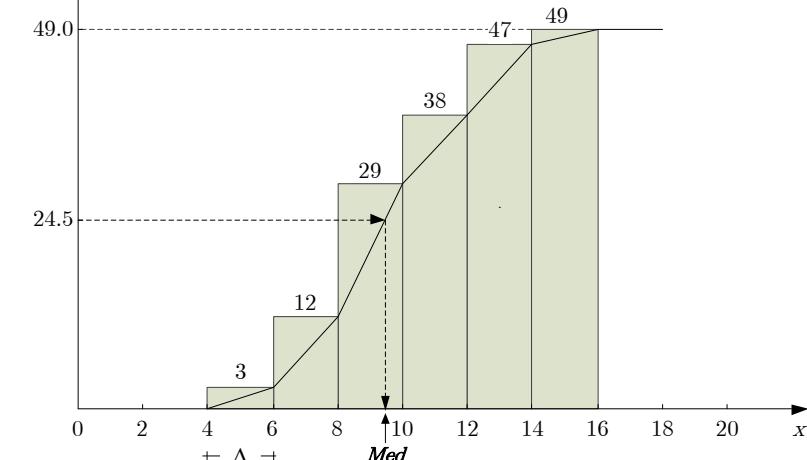
$\Delta$  - amplitude das classes



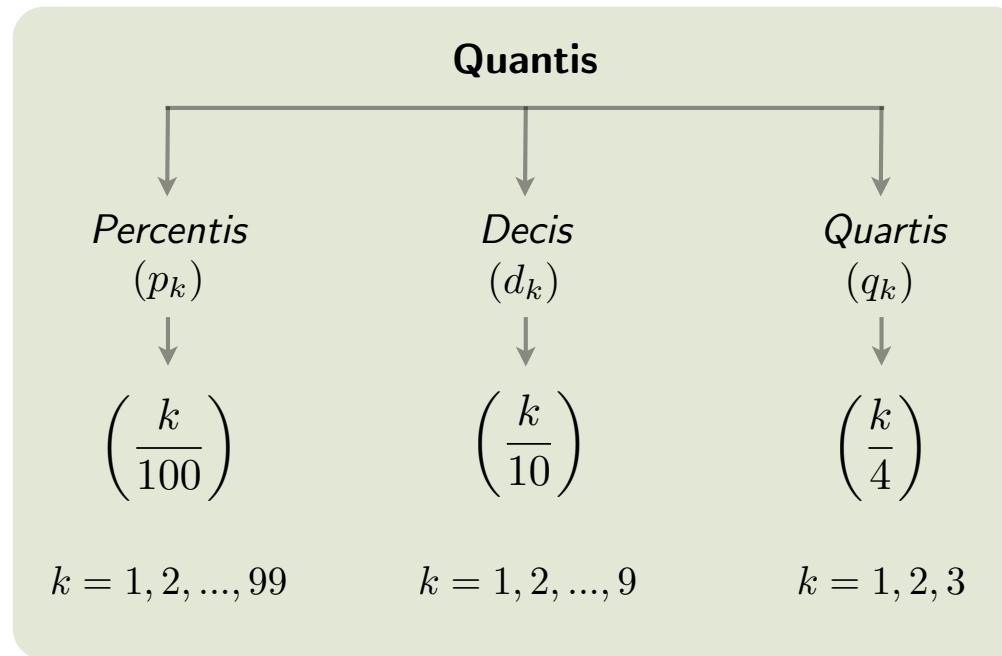
Cálculo do valor da mediana da classificação obtida pelos 49 alunos no exame de Estatística

$$Med = LI + \frac{n/2 - F'_{i-1}}{F'_i - F'_{i-1}} \times \Delta =$$

$$= 8 + \frac{49/2 - 12}{29 - 12} \times 2 = 9.471$$



#### 1.2.2.4. Quantis



- ▶  $Med = p_{50} = d_5 = q_2$
- ▶  $p_{10} = d_1, p_{20} = d_2, \dots, p_{90} = d_9$
- ▶  $p_{25} = q_1, p_{50} = q_2, \dots, p_{75} = q_3$

### 1.2.3. Estatísticas de dispersão

#### 1.2.3.1. Amplitude do Intervalo de Variação      (estatística básica de dispersão)

$$A = \max \{x_1, x_2, \dots, x_n\} - \min \{x_1, x_2, \dots, x_n\}$$

#### 1.2.3.2. Variância amostral      ( $s^2$ )

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{dados não agrupados})$$

$$s^2 = \frac{n}{n-1} \sum_{k=1}^K f_k \times (x_k - \bar{x})^2 \quad (\text{dados discretos agrupados})$$

$$s^2 = \frac{n}{n-1} \sum_{k=1}^K f_k \times (M_k - \bar{x})^2 \quad (\text{dados contínuos agrupados})$$

### 1.4.3.3 Desvio Padrão (s)

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

### 1.4.4. Outras Estatísticas

#### 1.4.4.1 Momentos

**Momentos na Origem de ordem 1, 2, ...**

$$m'_i = \frac{1}{n} \sum_{j=1}^n x_j^i$$

*dados não agrupados*

$$m'_i = \sum_{k=1}^K f_k \times x_k^i$$

*dados discretos agrupados*

$$m'_i \approx \sum_{k=1}^K f_k \times M_k^i$$

*dados contínuos agrupados*

## Momentos Centrados de ordem 1, 2, ...

$$m_i = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^i \quad \text{dados não agrupados}$$

$$m_i = \sum_{k=1}^K f_k \times (x_k - \bar{x})^i \quad \text{dados discretos agrupados}$$

$$m_i \approx \sum_{k=1}^K f_k \times (M_k - \bar{x})^i \quad \text{dados contínuos agrupados}$$

### ***Estimadores não enviesados dos momentos centrados***

$$k_2 = s^2 = \frac{n}{n-1} \times m_2 \quad 2^{\text{a}} \text{ ordem}$$

$$k_3 = \frac{n^2}{(n-1)(n-2)} \times m_3 \quad 3^{\text{a}} \text{ ordem}$$

$$k_4 = \frac{n \times (n^2 - 2n + 3)}{(n-1)(n-2)(n-3)} \times m_4 - \frac{3n \times (2n-3)}{(n-1)(n-2)(n-3)} \times m_2^2 \quad 4^{\text{a}} \text{ ordem}$$

#### 1.4.4.2. Coeficiente de Assimetria

$$CA = \frac{m_3}{m_2^{3/2}} \quad \text{enviesado}$$

$$CA' = \frac{k_3}{k_2^{3/2}} \quad \text{enviesado (menos que o anterior)}$$

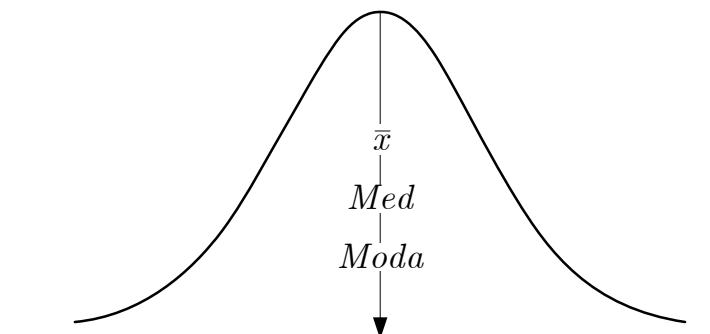
$$CA'' = \frac{\sqrt{n \times (n - 1)}}{n - 2} \times CA \quad \text{não enviesado para populações normais}$$

$$\left\{ \begin{array}{ll} \text{Distribuição simétrica} & \rightarrow C_A = 0 \\ \text{Distribuição assimétrica} & \left\{ \begin{array}{ll} \text{assimétrica à esquerda} & \rightarrow C_A < 0 \\ \text{assimétrica à direita} & \rightarrow C_A > 0 \end{array} \right. \end{array} \right.$$

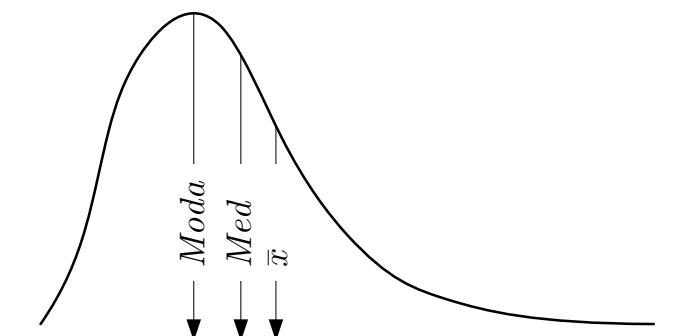
## Posição da Média, Moda e Mediana

DGI

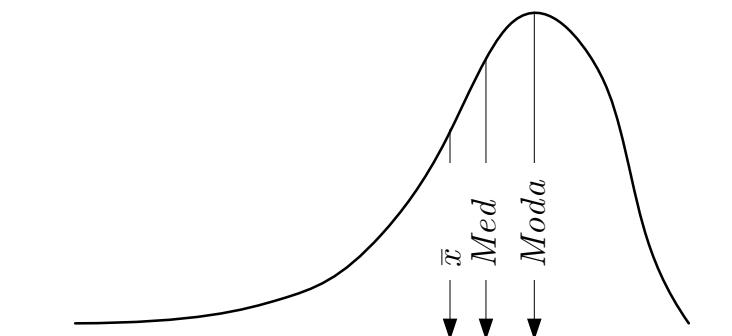
2019



(distribuição simétrica)



(distribuição assimétrica à direita)



(distribuição assimétrica à esquerda)

**1.4.4.3. Coeficiente de Achatamento ou de Kurtosis**

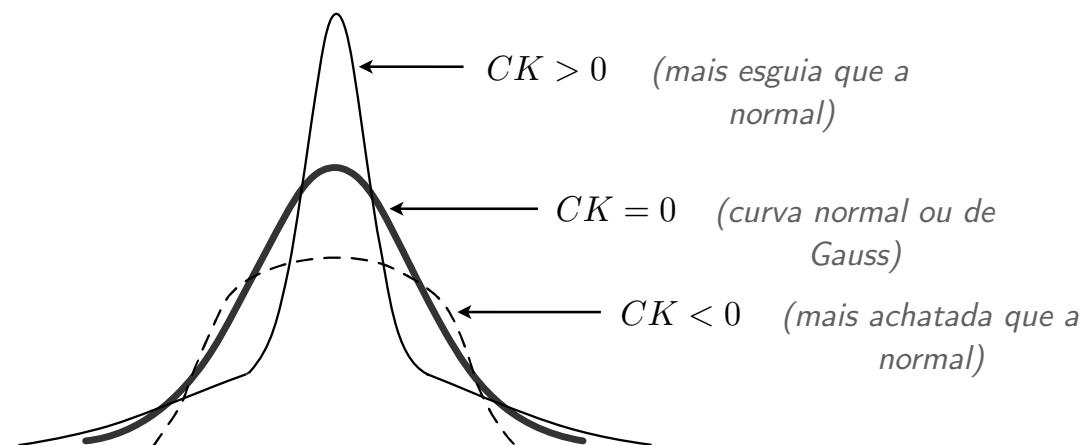
$$CK = \frac{m_4}{m_2^2} - 3$$

enviesado

$$CK' = \frac{k_4}{k_2^2} - 3$$

enviesado (menos que o anterior)

$$CK'' = \frac{n-1}{(n-2)(n-3)} \times [(n+1)CK + 6] \quad \text{não enviesado para populações normais}$$



## Exemplo

Calcule os valores dos coeficientes de assimetria e de achatamento relativos à classificação obtida pelos 49 alunos no exame de Estatística.

DGI

2019

► *Considerando os dados não agrupados:*

$$m_2 = 6.52980$$

$$m_3 = 1.86658$$

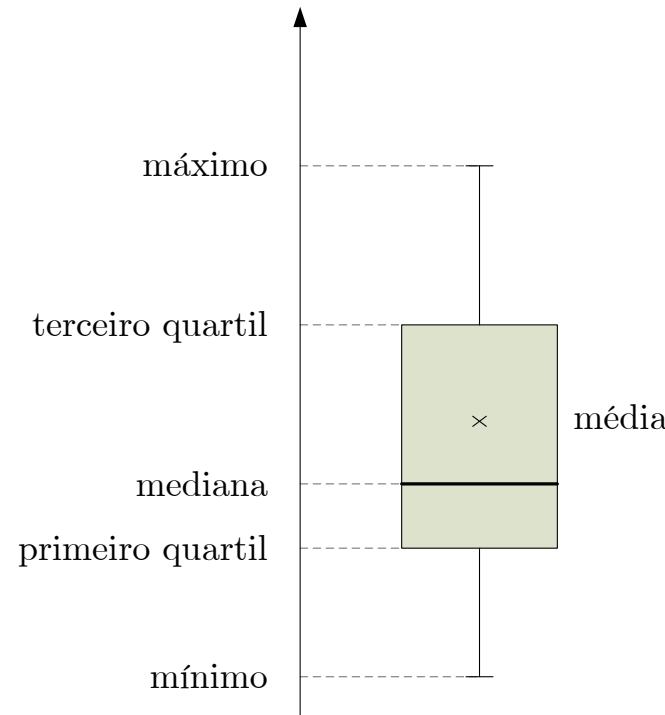
$$m_4 = 95.39174$$

$$CA = 0.11187$$

$$CK = -0.7628$$

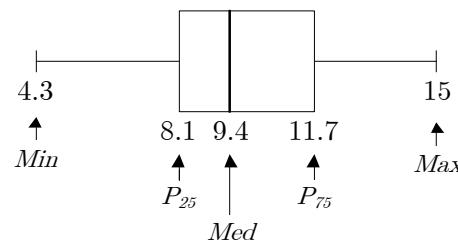
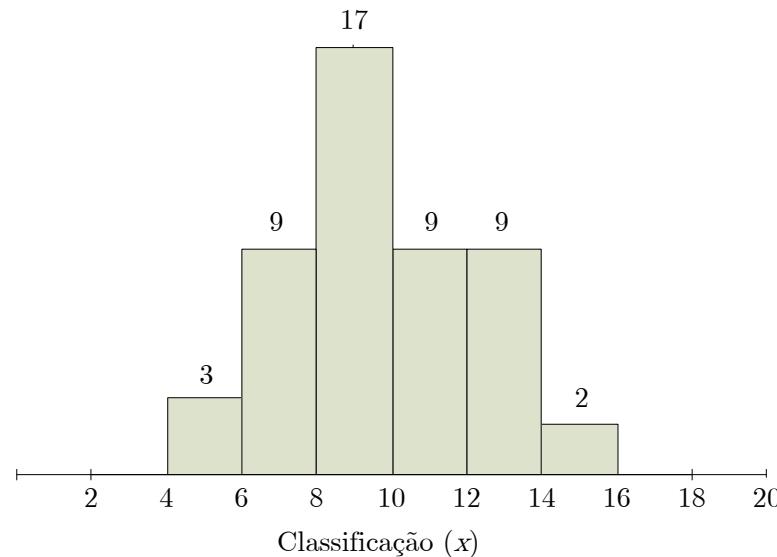
## 1.4.5. Representação Gráfica de Estatísticas

### 1.4.5.1. Box & Whisker Plots



## Exemplo

Construção do *box-plot* relativo à classificação obtida pelos 49 alunos no exame de estatística



## 1.5. Caracterização de Amostras Bivariadas

Uma amostra bivariada é constituída por pares ordenados  $(x,y)$ , sendo  $x$  e  $y$ , atributos de um mesmo objecto

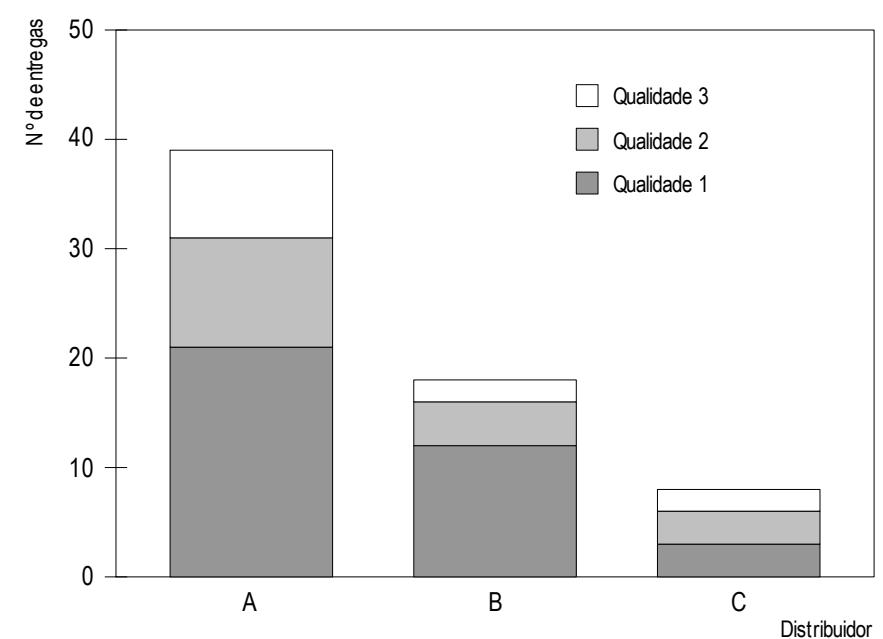
### 1.5.1. Dados Qualitativos

As formas mais usuais de caracterizar amostras bivariadas envolvem o recurso a **tabelas de informação cruzada e a diagramas de barras sobrepostas.**

## Tabela de Informação Cruzada

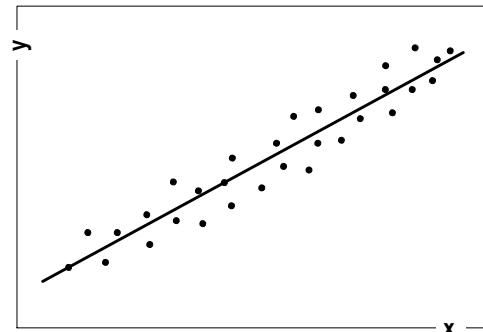
Distribuidor	Qualidade das Entregas			Total da Linha
	1	2	3	
A	21	10	8	39
	32.3%	15.4%	12.3%	60.0%
	53.8%	25.6%	20.5%	100.0%
	58.3%	58.8%	66.7%	-
B	12	4	2	18
	18.5%	6.2%	3.1%	27.7%
	66.7%	22.2%	11.1%	100.0%
	33.3%	23.5%	16.7%	-
C	3	3	2	8
	4.6%	4.6%	3.1%	12.3%
	37.5%	37.5%	25.0%	100.0%
	8.3%	17.6%	16.7%	-
Total da Coluna	36	17	12	65
	55.4%	26.2%	18.5%	100.0%
	100.0%	100.0%	100.0%	-

## Diagrama de Barras Sobrepostas

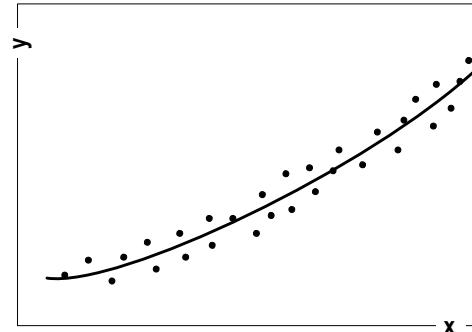


### 1.5.2. Dados Quantitativos

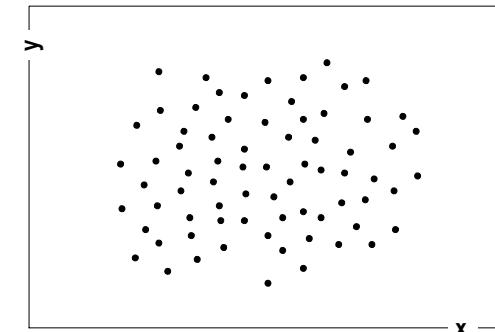
- ▶ A melhor forma de caracterizar a relação entre as duas variáveis é a sua representação conjunta num sistema de eixos ortogonais - diagrama de dispersão
- ▶ A construção do diagrama de dispersão apresenta uma dupla função: Ajuda a determinar se existe alguma relação entre as variáveis e, caso exista, permite identificar a equação mais adequada para a descrever.



(relação linear)



(relação não linear)



(ausência de relação)

### 1.5.2.1. Modelo de Regressão Linear

$$y_i = f(\theta_1, \theta_2, \dots, \theta_k) = f(a, b) = a + b \cdot x_i$$

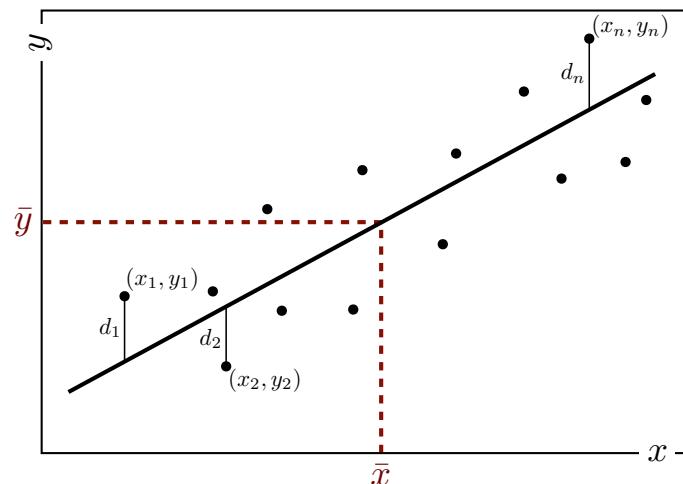
DGI

2019

sendo:  $y$  - variável dependente,  $x$  - variável independente ,  $a$  - ordenada na origem

### Método dos Mínimos Quadrados

- Ajustamento de uma *relação linear* aos dados observados que *minimiza o somatório do quadrado das distâncias entre os valores observados e o modelo ajustado.*



$$d_1^2 + d_2^2 + \dots + d_n^2 = \text{mínimo}$$

$$SEQ = SEQ(a, b) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N [y_i - (a + b \cdot x_i)]^2$$

**minimizar:**

$$\left\{ \begin{array}{l} \frac{\partial SEQ(a, b)}{\partial a} = 0 \\ \frac{\partial SEQ(a, b)}{\partial b} = 0 \end{array} \right. \rightarrow \begin{array}{l} a = \frac{\frac{1}{N} \sum_{i=1}^N x_i^2 \cdot \sum_{i=1}^N y_i - \frac{1}{N} \sum_{i=1}^N x_i \cdot \sum_{i=1}^N x_i \cdot y_i}{\sum_{i=1}^N x_i^2 - \frac{1}{N} \cdot \left( \sum_{i=1}^N x_i \right)^2} \\ b = \frac{\sum_{i=1}^N x_i \cdot y_i - \frac{1}{N} \sum_{i=1}^N x_i \cdot \sum_{i=1}^N y_i}{\sum_{i=1}^N x_i^2 - \frac{1}{N} \cdot \left( \sum_{i=1}^N x_i \right)^2} \end{array}$$

## Coeficiente de Correlação Amostral (r)

- Medida do **grau de relacionamento** linear entre os dados de uma amostra bivariada.

DGI

2019

$$r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\sum_{i=1}^N x_i \cdot y_i - \frac{1}{N} \sum_{i=1}^N x_i \cdot \sum_{i=1}^N y_i}{\sqrt{\left[ \sum_{i=1}^N x_i^2 - \frac{1}{N} \left( \sum_{i=1}^N x_i \right)^2 \right] \cdot \left[ \sum_{i=1}^N y_i^2 - \frac{1}{N} \left( \sum_{i=1}^N y_i \right)^2 \right]}}$$

sendo:  $S_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})$

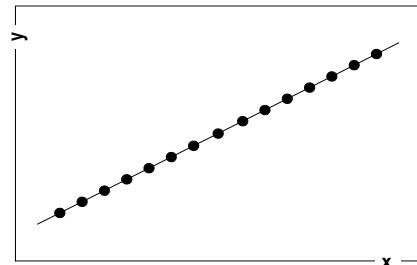
$$S^2_x = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

$$S^2_y = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2$$

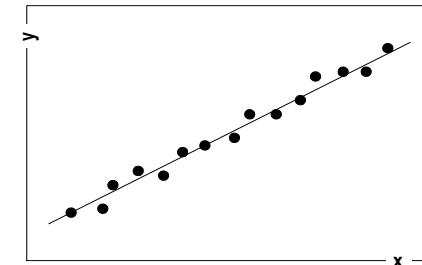
## Variação do Coeficiente de Correlação Amostral

DGI

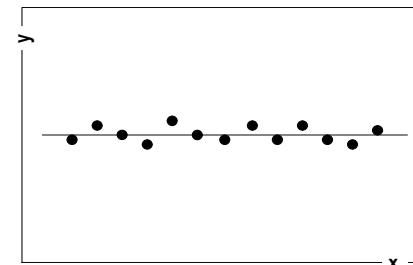
2019



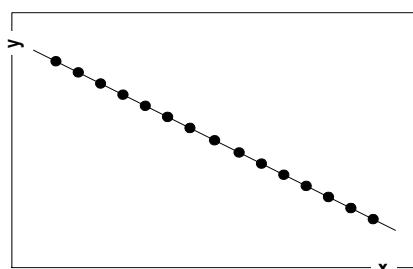
$$r = 1$$



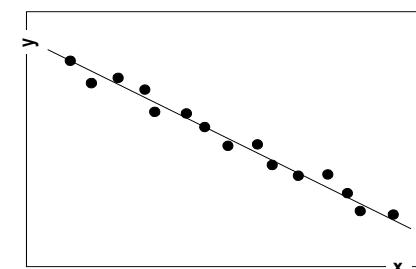
$$0 < r < 1$$



$$r = 0$$



$$r = -1$$



$$0 > r > -1$$

## ***Coeficiente de Determinação ( $r^2$ )***

- Traduz a proporção da variação da variável dependente ( $Y$ ) que é explicada pela variação da variável independente ( $X$ ), através do modelo de regressão.
  - Variação do Coeficiente de Determinação

$$0 \leq r^2 \leq 1$$

## Exemplo

Os dados representados a seguir referem-se ao volume de vendas mensais, expresso em unidades monetárias, de um artigo comercializado por uma empresa. Os valores representados referem-se aos últimos 12 meses.

mês	1	2	3	4	5	6	7	8	9	10	11	12
vol. vendas	3.7	4.1	5.9	6.6	6.4	7.2	8	8.8	9.4	9.3	9.7	9.9

Com base nos dados disponíveis determine:

- A média e a variância amostral do volume de vendas mensal.
- A correlação amostral entre os meses e o volume de vendas e, ainda, a previsão do volume de vendas para o próximo mês.

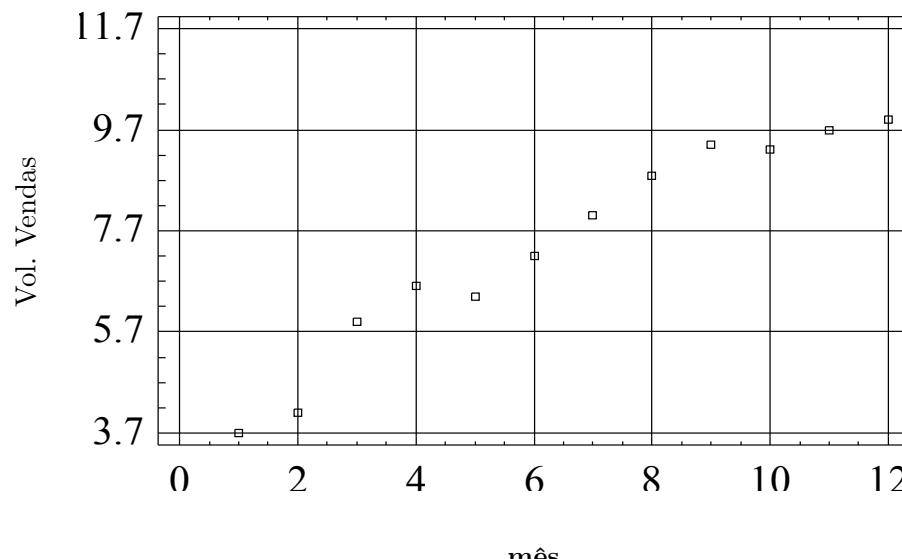
## Resolução

a)  $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{3.7+4.1+\dots+9.9}{12} = 7.4167$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{12-1} \cdot \left( (3.7 - 7.4167)^2 + \dots + (9.9 - 7.4167)^2 \right) = 4.5433$$

b)

Diagrama de Dispersão



### Parâmetros do modelo:

$$a = \frac{\frac{1}{N} \sum_{i=1}^N x_i^2 \cdot \sum_{i=1}^N y_i - \frac{1}{N} \sum_{i=1}^N x_i \cdot \sum_{i=1}^N x_i \cdot y_i}{\sum_{i=1}^N x_i^2 - \frac{1}{N} \cdot \left( \sum_{i=1}^N x_i \right)^2} = 3.6803$$

$$b = \frac{\sum_{i=1}^N x_i \cdot y_i - \frac{1}{N} \sum_{i=1}^N x_i \cdot \sum_{i=1}^N y_i}{\sum_{i=1}^N x_i^2 - \frac{1}{N} \cdot \left( \sum_{i=1}^N x_i \right)^2} = 0.5748$$



$$y = 0.5748 \cdot x + 3.6803$$

$$r = \frac{\sum_{i=1}^N x_i \cdot y_i - \frac{1}{N} \sum_{i=1}^N x_i \cdot \sum_{i=1}^N y_i}{\sqrt{\left[ \sum_{i=1}^N x_i^2 - \frac{1}{N} \left( \sum_{i=1}^N x_i \right)^2 \right] \cdot \left[ \sum_{i=1}^N y_i^2 - \frac{1}{N} \left( \sum_{i=1}^N y_i \right)^2 \right]}} = 0.9723 \rightarrow r^2 = 0.9455$$

Previsão para o 13º mês ( $x = 13$ ):  $y = 0.5748 \times 13 + 3.6803 = 11.153$

