ESTATÍSTICA DESCRITIVA

- Organização
- Descrição
- Quantificação de variabilidade
- Identificação de valores típicos e atípicos

- **Elementos básicos:**
  - Tabelas
  - Gráficos
  - Resumos numéricos

CONCEITOS BÁSICOS

- **Variável**

Quantificação ou categorização do fenômeno de interesse (na Medicina chamado de parâmetro)

Inquérito epidemiológico com as perguntas:

Pergunta	Variável
Qual é a sua idade?	Idade
Qual o número de pessoas na família?	Tamanho da família
Qual é a renda total de sua família?	Faixa de renda
Qual é o seu estado civil?	Estado civil
Você tem emprego fixo?	Emprego

Tipos de Variáveis

- Facilita o tratamento estatístico classificar variáveis em: **Categóricas e Quantitativas**

- **Variáveis Categóricas**
  - **Nominais:** Emprego, Estado civil, Tipo sanguíneo
  - **Ordinais:** Faixa de renda, Grau de Escolaridade

Tipos de Variáveis

- **Variáveis Quantitativas**
  - **Discretas:** Tamanho da família, Número de bactérias num volume de urina, Número de consultas no mês, Número de batimentos cardíacos por minuto
  - **Contínuas:** Idade, pressão sanguínea, peso, altura

V. Discreta x V. Categ. Ordinal

- A ordenação tem significado diferente:
  - **Número de crianças** (0, 1, 2, 3, 4): 4 crianças corresponde ao dobro de 2 crianças, e uma família com 4 crianças tem uma criança a mais do que uma família com 3, que por sua vez tem uma criança a mais do que uma família com 2 crianças.
  - **Estadiamento de câncer de mama** (I, II, III e IV): Não se pode dizer que IV é duas vezes pior do que II, ou que a diferença entre I e II é equivalente à entre III e IV.

DADOS BRUTOS

- Obtidos diretamente da pesquisa
- Ainda não sofreram processo de análise ou síntese
- Apresentados em tabelas mas omitidos das publicações por questão de espaço
- O conjunto de dados constitui uma ***amostra***. O tamanho da amostra é geralmente denotado por ***n***.

Exemplo: Teor de gordura fecal em crianças

- Utilidade diagnóstica do teor de gordura fecal inquestionável mas até 1984 não existia um padrão de referência para crianças brasileiras.
- Prof. Francisco Penna (titular de Pediatria da UFMG) examinou 43 crianças sadias

Tabela: Teor de gordura fecal (g/24 hs)

3,7	1,6	2,5	3,0	3,9	1,9	3,8	1,5	1,1
1,8	1,4	2,7	3,3	3,2	2,3	2,3	2,3	2,4
0,8	3,1	1,8	1,0	2,0	2,0	2,9	3,2	1,9
1,6	2,9	2,0	1,0	2,7	3,0	1,3	1,5	4,6
2,4	2,1	1,3	2,7	2,1	2,8	1,9		

- Note a grande variação dos resultados!
- Como definir um padrão de referência?

Exemplo 3.2: Nível de colesterol

- 1948, cidade de Framingham selecionada para um estudo prospectivo
- **Objetivo:** verificar como hábitos de vida influenciam o risco de desenvolvimento doenças cardíacas.
- **Resultado:** Necessidade de controle do nível de colesterol.

Tabela: taxa de colesterol (mg/dL) em 1952

278	182	247	227	277	194	196	276	244	192
118	219	255	201		209	219	228	209	209
171	213	233	226	209	200	200	363	209	200
179	167	192	277	317	146	217	292	217	255
212	233	250	243	150	209	184	199	250	479
175	194	221	233		184	217	150	167	265
242	180	255	170	209	161	196	165	234	179
248	184	291	185	242	276	243	229	242	250

- Observando a tabela o que apreendemos sobre o nível do colesterol à época do exame?
- Como saber o valor em torno do qual as medidas estão agrupadas, a forma da distribuição e a extensão da variabilidade?

Organização e apresentação de dados

- Para uma variável ou para o cruzamento de variáveis
  - Tabelas de frequências
  - Gráficos

Tabelas de frequências

- Sintetiza os dados
- Consiste na construção de uma tabela a partir dos dados brutos com a frequência de cada observação.
- A partir das tabelas são construídos os gráficos.

Exemplo 3.5: Tentativas de suicídio

- Estudo retrospectivo (Fernandes et al., 1995)
- Tentativas de suicídio por intoxicação aguda registradas no Centro de Assistência Toxicológica do Hospital de Base de São Paulo.
- Período de 01/92 a 02/93
- 302 casos
- 27% do total de atendimentos no período
- 67% das tentativas de suicídio do sexo feminino

Tabela 3.3: Distribuição de profissões entre pacientes potencialmente suicidas

Profissão	Frequência	Proporção
Serviços Gerais*	75	0,248
Doméstica**	55	0,182
Do Lar	53	0,175
Indeterminada	29	0,096
Emprego especializado***	23	0,076
Menor	20	0,066
Desempregado	15	0,050
Estudante	14	0,046
Lavrador	12	0,040
Autônomo	4	0,013
Aposentado	2	0,007
Total	302	1

* garçom, encanador, pedreiro, frentista, operário, padeiro, açougueiro, borracheiro, etc.

** copeira, faxineira, costureira, bordadeira

*** enfermeiro, modelo, protético, escrivão, professor, digitador, vendedor

Tabela 3.4: Distribuição de tentativas de suicídio segundo faixa etária

Idade (anos)	Frequência	
	Absoluta	Relativa
10|-20	57	18,87
20|-30	113	37,42
30|-40	59	19,54
40|-50	32	10,60
50|-60	19	6,29
60|-70	7	2,32
≥70	2	0,66
Indeterminada	13	4,30
Total	302	100

Tabela: taxa de colesterol (mg/dL) em 1952

278	182	247	227	277	194	196	276	244	192
118	219	255	201		209	219	228	209	209
171	213	233	226	209	200	200	363	209	200
179	167	192	277	317	146	217	292	217	255
212	233	250	243	150	209	184	199	250	479
175	194	221	233		184	217	150	167	265
242	180	255	170	209	161	196	165	234	179
248	184	291	185	242	276	243	229	242	250

- Observando a tabela o que apreendemos sobre o nível do colesterol à época do exame?
- Como saber o valor em torno do qual as medidas estão agrupadas, a forma da distribuição e a extensão da variabilidade?

Exemplo 3.2: Taxas de Colesterol (cont.)

- n=78, Min=118, Max=479
- Número de classes=1+log(n,base=2)=7,28≈8
- Tamanho de classe=(479-118)/8=45,125≈50

Tabela 3.10: Distribuição do nível de colesterol

Nível de	Frequência absoluta		Frequência relativa	
Colesterol	simples	acumulada	simples	acumulada
100|-150	2	2	0,03	0,03
150|-200	24	26	0,31	0,34
200|-250	35	61	0,45	0,79
250|-300	14	75	0,18	0,97
300|-350	1	76	0,01	0,98
350|-400	1	77	0,01	0,99
400|-450	0	77	0	0,99
450|-500	1	78	0,01	1
Total	78	-	1	-

Etapas para construção de tabelas de frequências para dados agrupados

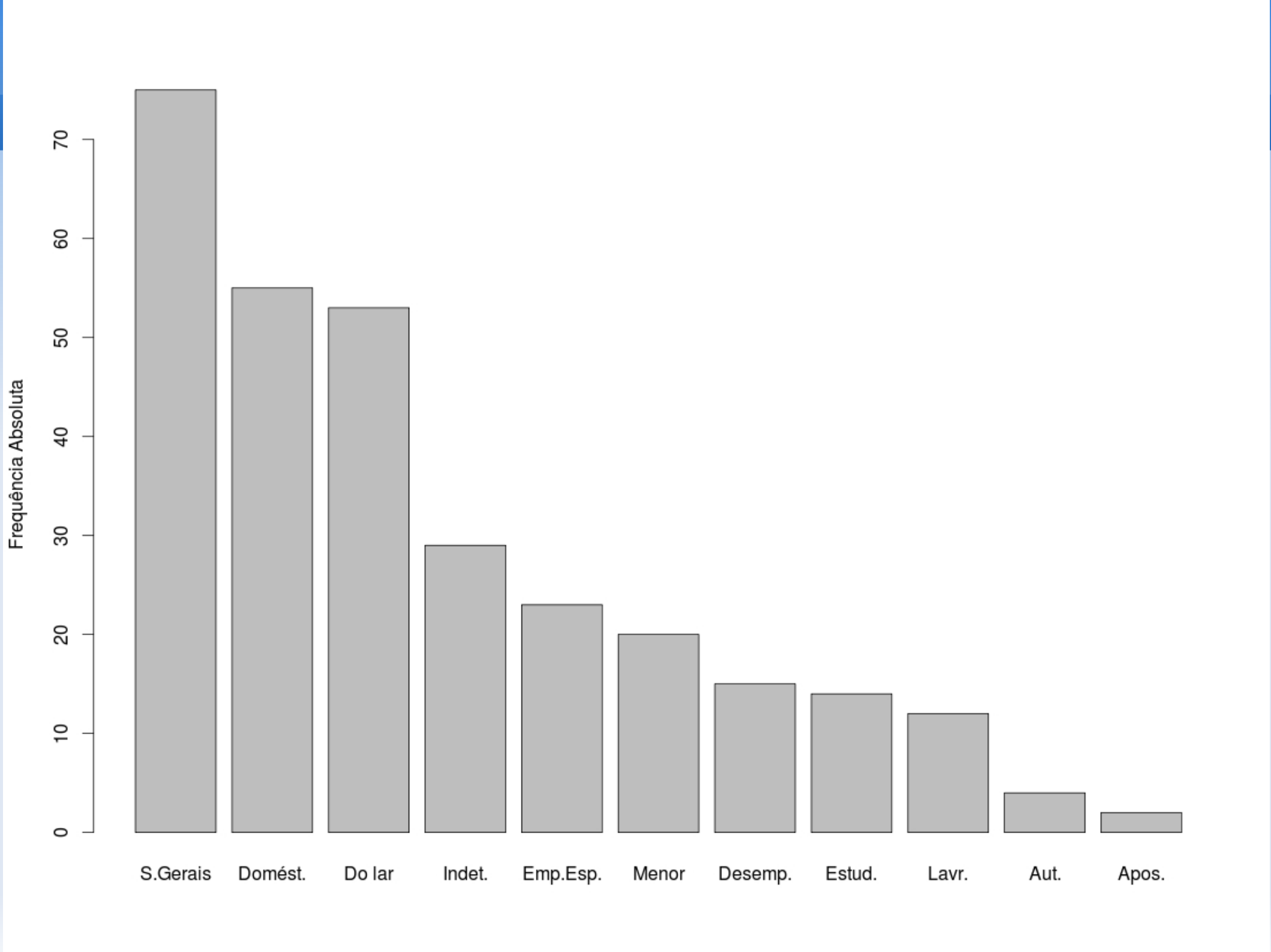
1. Encontrar o menor e o maior valores (mínimo e máximo) do conjunto de dados
2. Escolher número de classes (de igual amplitude), que englobem todos os dados sem superposição de intervalos.
3. Contar o número de elementos em cada classe (este número é a frequência absoluta)
4. Calcular a frequência relativa em cada classe

GRÁFICOS

- Diagrama de barras
- Histograma
- Ogiva
- Gráfico de linhas
- Diagrama de pontos
- Diagrama de dispersão

Representação gráfica para variáveis categóricas

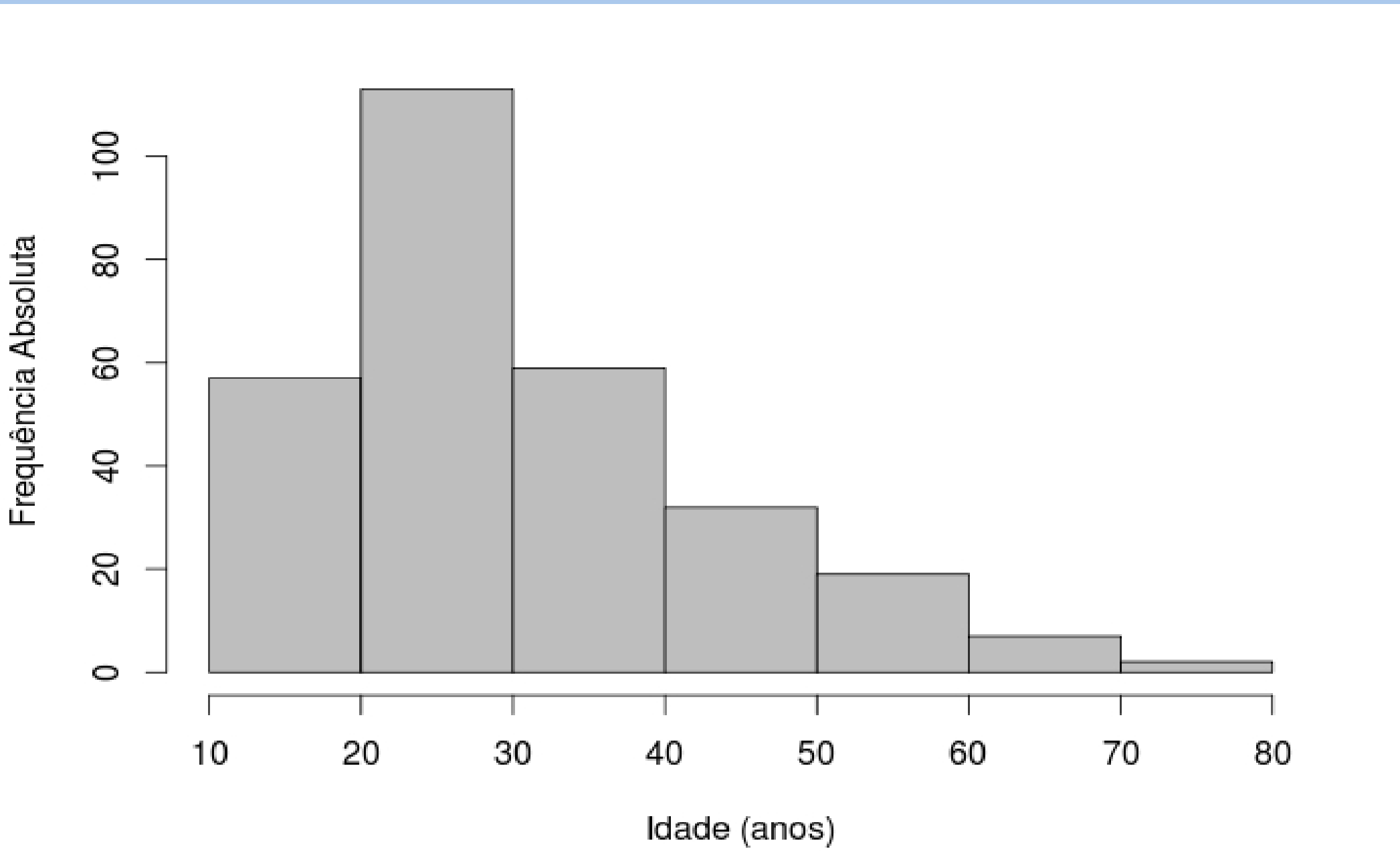
- Diagrama de barras
- Exemplo 3.5: Distribuição de profissões entre pacientes potencialmente suicídas (cont.)


Frequência Absoluta
0
10
20
30
40
50
60
70
S.Gerais
Domést.
Do lar
Indet.
Emp.Esp.
Menor
Desemp.
Estud.
Lavr.
Aut.
Apos.

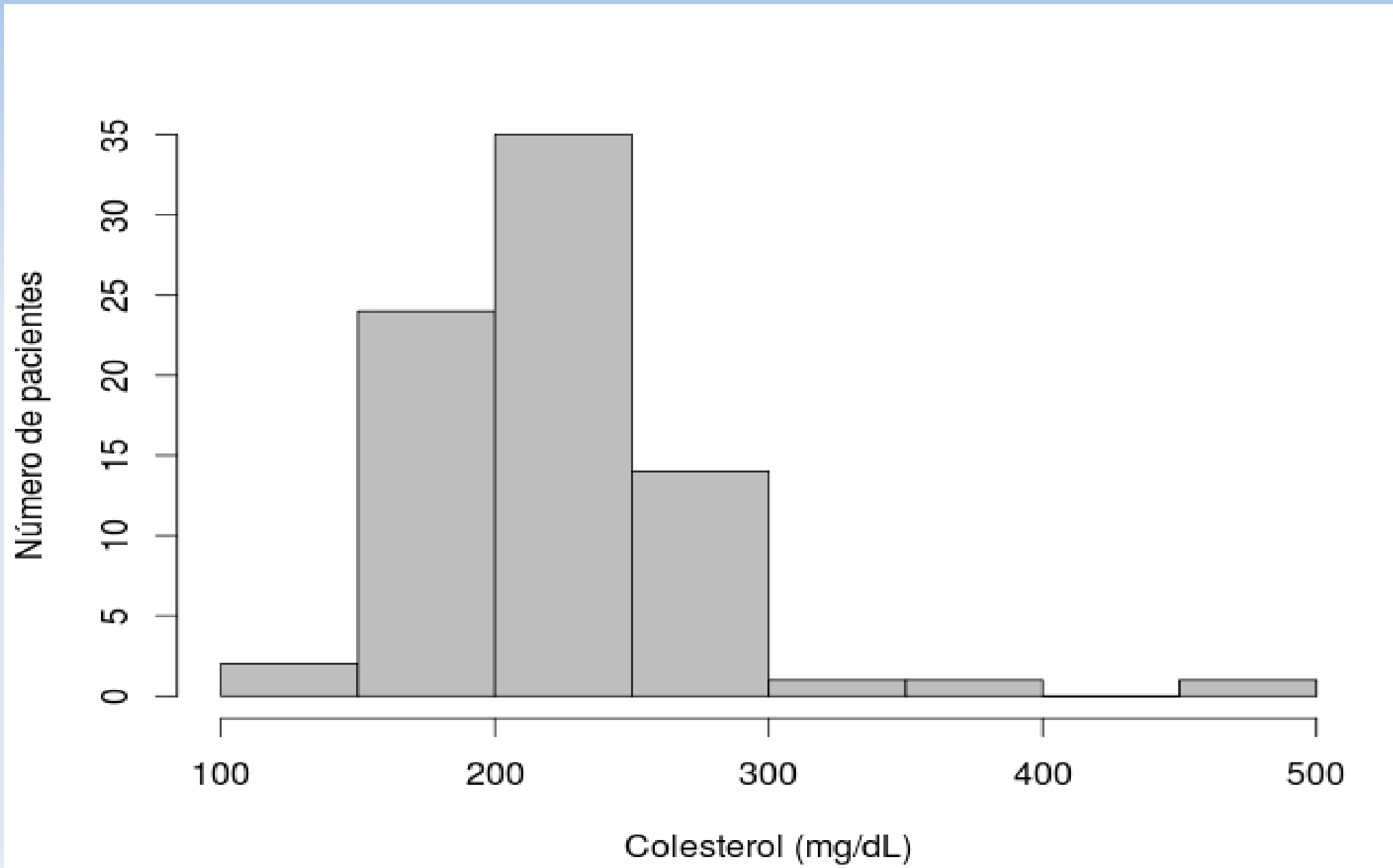
Representação gráfica de variáveis quantitativas

- Histograma
  - Serve para visualizar a forma da distribuição da variável estudada.

Exemplo 3.5: Distribuição das tentativas de suicídio segundo faixa etária



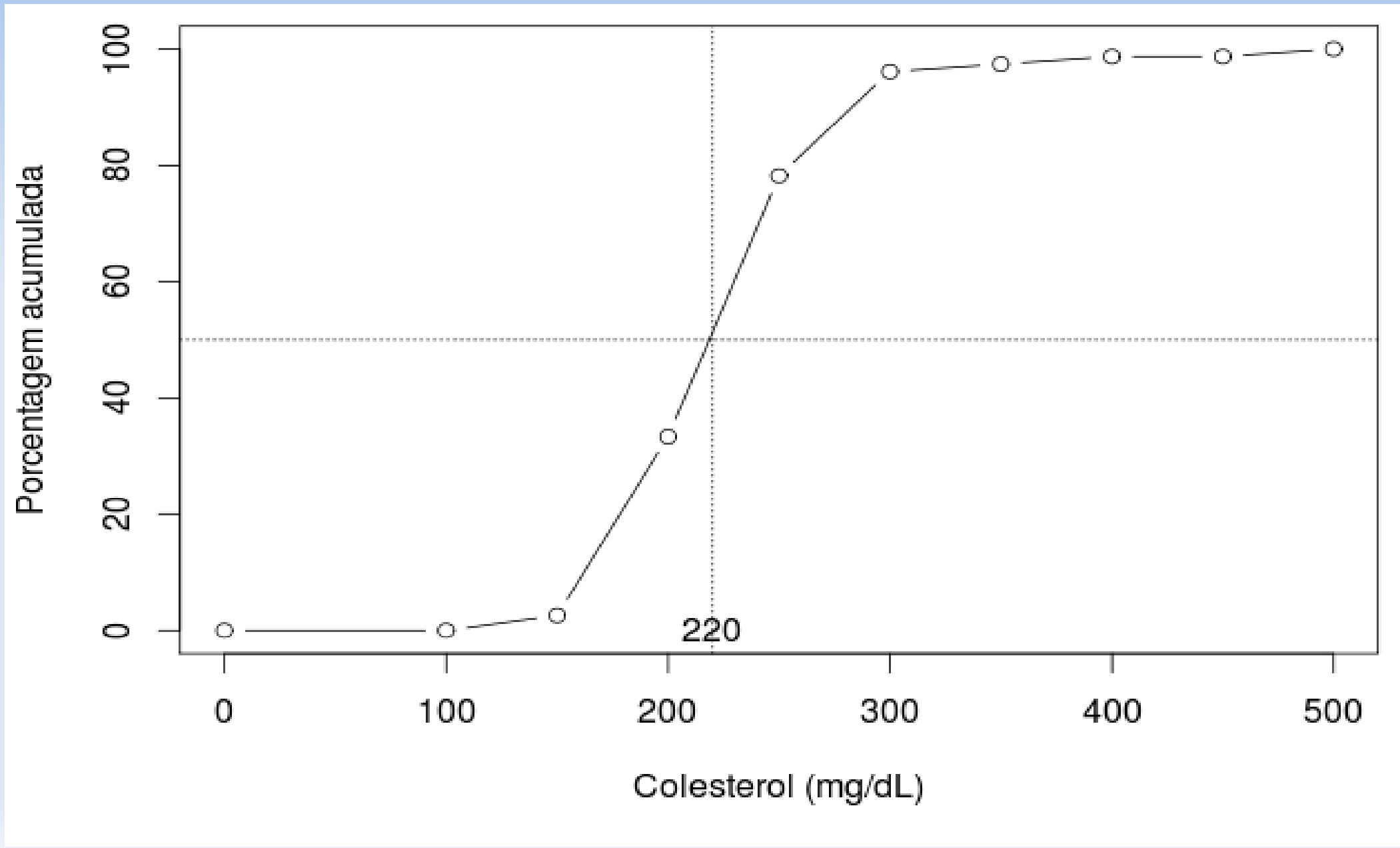
Exemplo 3.2: Distribuição do nível de colesterol



Ogiva

- Gráfico de percentuais acumulados
- Através da ogiva podemos estimar percentis da distribuição, isto é, o valor que é precedido por uma porcentagem pré-estabelecida.
- Exemplo: estimar o valor da variável abaixo do qual se tem 50% dos indivíduos.

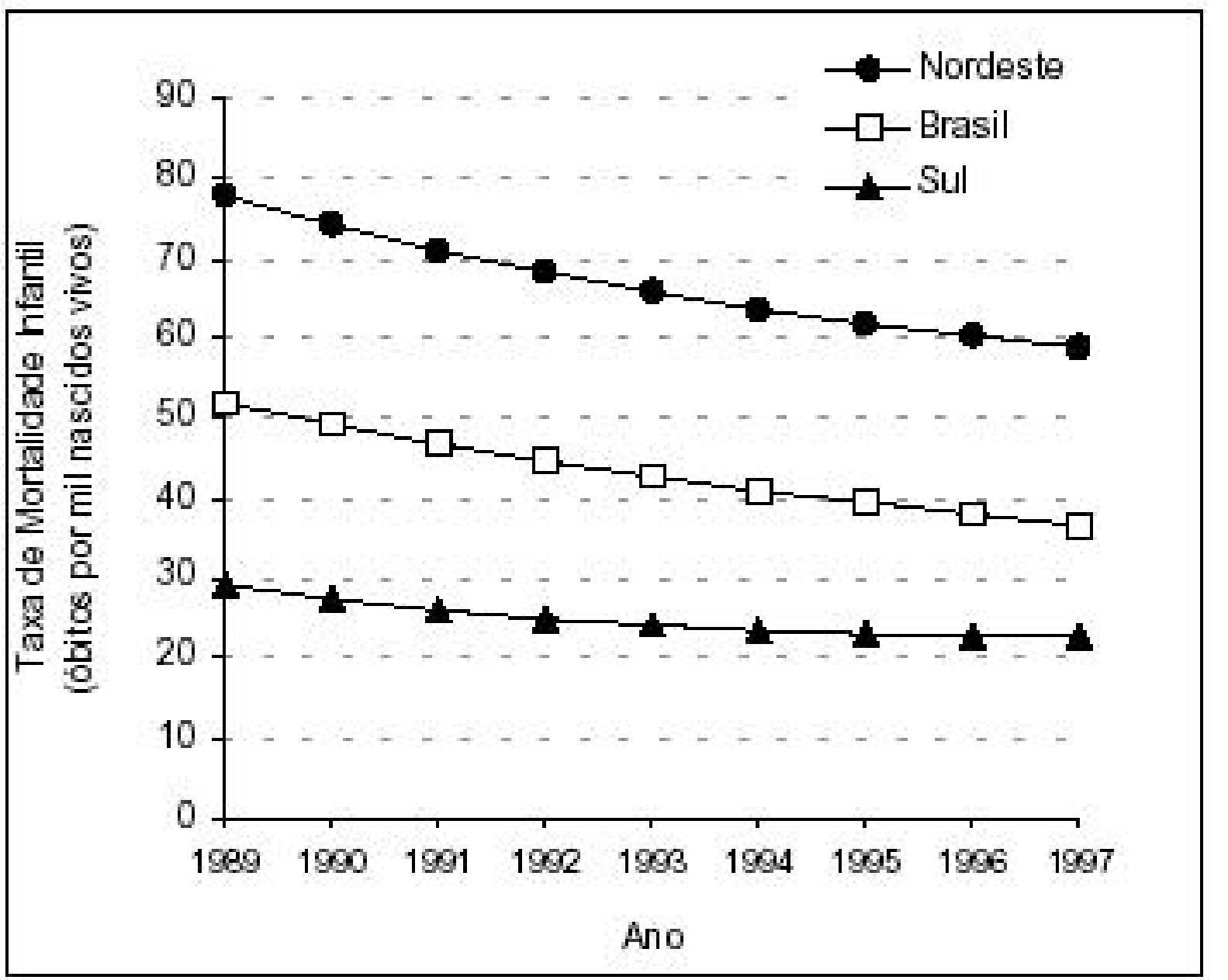
Ogiva das taxas de colesterol



Representação gráfica de dados temporais

- Dados coletados ao longo do tempo são comuns em pesquisas médicas
- Diagrama de barras para períodos agrupados (ex: menos de 1 ano, 1 a 5 anos, 5 a 10 anos)
- Gráfico de linhas é o mais apropriado
  - Eixo horizontal: escala temporal
  - Eixo vertical: variável de interesse
- Permite constatar tendências e identificar eventos extremos

Representação gráfica de dados temporais



Coleta de dados

- Há várias maneiras de se coletar dados, dependendo do tipo de estudo
- **Experimento com cobaias:** medida direta das variáveis de interesse
- **Inquérito:** questionário é o instrumento de medida mais utilizado
- **Pesquisas clínicas:** formulário, prontuário do paciente, ou ficha preenchida na anamnese.
- Vários cuidados devem ser tomados na elaboração e utilização de um instrumento de pesquisa (refs na pág.64 do livro texto)

Banco de dados

- Uma linha para cada indivíduo
- Uma coluna para cada variável observada
- Para variáveis categóricas:
  - Criar códigos para cada categoria
- Para variáveis contínuas:
  - Entrar com os dados originais e não os codificados para classes de interesse, pois pode haver mudança nas classes de interesse durante a análise
- Para dados omissos: usar código que facilmente identifique esse tipo de dado (Ex: 999 para pressão arterial)

Exemplo 3.20: Tentativas de suicídio (cont.)

Dicionário das variáveis:

- Sexo: 0 para masculino e 1 para feminino
- Profissão: 1-Serviços Gerais, 2-Doméstica, 3-Do lar, 4-Inderteminado, 5-Emprego Especializado, 6-Menor, 7-Desempregado, 8-Estudante, 9-Lavrador, 10-Autônomo, 11-Aposentado
- Idade: anos

Indivíduo	Sexo	Profissão	Idade
1	0	1	25
2	1	2	48
...	...	...	...
302	1	8	13

RESUMOS NUMÉRICOS

- MEDIDAS DE TENDÊNCIA CENTRAL
  - Moda
  - Média
  - Mediana
- MEDIDAS DE DISPERSÃO OU VARIABILIDADE
  - Amplitude
  - Variância
  - Desvio-padrão
  - Coeficiente de variação
  - Escore padronizado

MEDIDAS DE TENDÊNCIA CENTRAL

- EXEMPLO: Os dados a seguir referem-se a um grupo de pacientes submetidos a um teste sorológico realizado no sangue.

paciente	sexo	tipo.sangue	idade	reação	tempo.de.reação
1	M	A	8	negativa	15,5
2	F	O	46	positiva	8,7
3	M	B	50	negativa	2,8
4	F	O	42	positiva	11,9
5	F	O	52	positiva	5
6	M	A	56	positiva	9,7
7	M	AB	42	negativa	13
8	M	B	38	negativa	7,1
9	F	A	48	negativa	11,1
10	M	A	58	negativa	5,7
11	M	A	11	positiva	6,3
12	M	O	46	positiva	15,1
13	F	O	35	negativa	10,7
14	F	B	56	negativa	11,7
15	F	B	19	negativa	13,3
16	F	AB	28	positiva	8,8
17	F	A	44	negativa	8,3
18	M	O	52	negativa	16,9
19	M	O	34	positiva	9,1
20	F	A	21	positiva	7,8
21	F	B	35	negativa	13,1
22	M	A	34	positiva	13,5
23	F	AB	50	positiva	15,4
24	F	A	46	negativa	10,8
25	M	B	45	negativa	11,2
26	M	AB	42	negativa	3,6
27	F	O	58	negativa	9,8
28	F	O	45	positiva	7,2
29	M	A	44	negativa	12,8
30	F	A	22	negativa	10,6

Moda

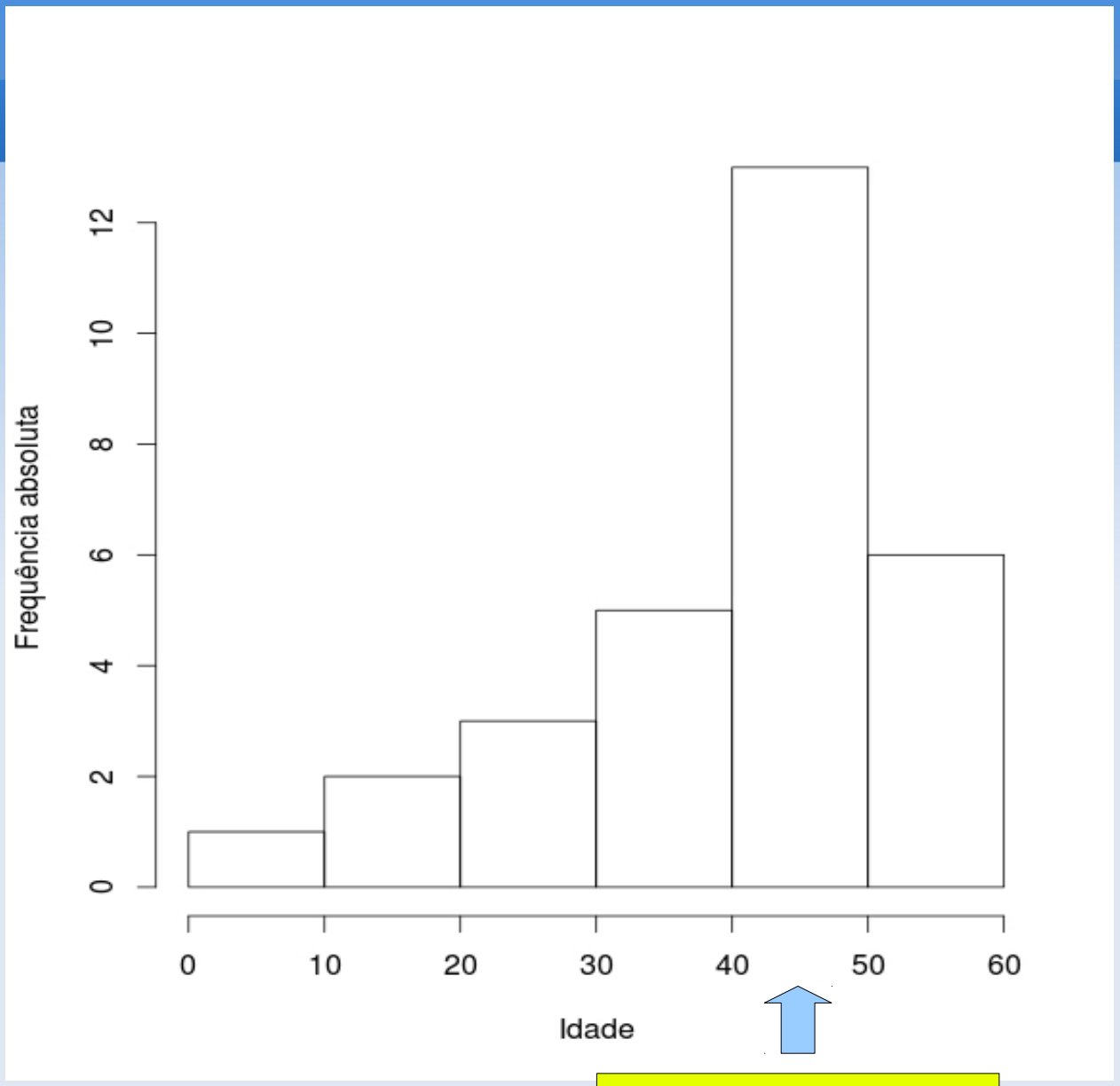
Característica ou valor que ocorre com maior freqüência.

Tabela - Distribuição dos pacientes quanto à tipagem sanguínea

tipo sangue	frequência
A	11
B	6
AB	4
O	9
Total	30

Moda: sangue do tipo A

Com dados quantitativos de natureza contínua, em geral, basta identificar a classe modal.

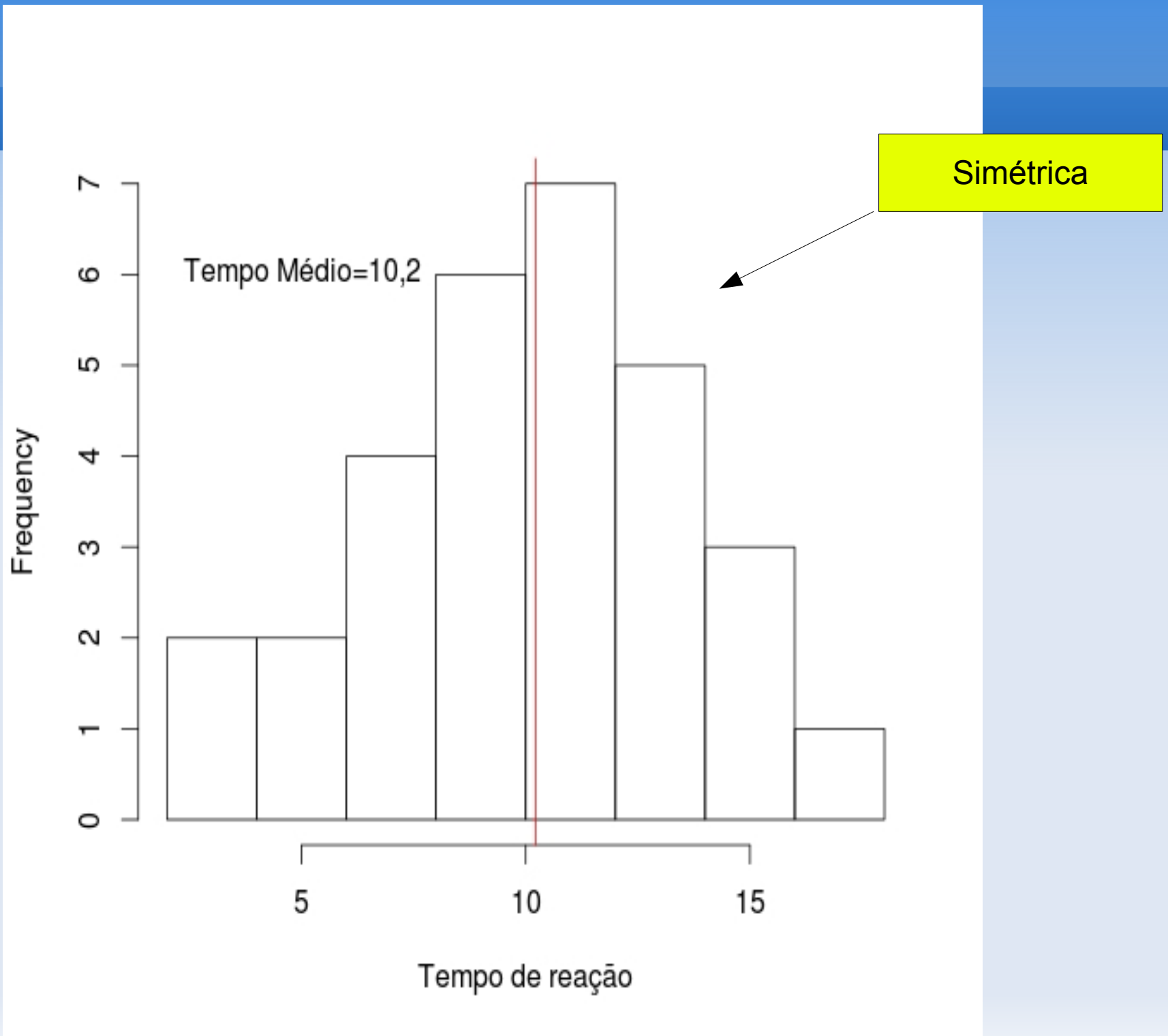


Classe modal

MEDIDAS DE TENDÊNCIA CENTRAL

- Média

$$\bar{x} = \frac{\sum x}{n}$$


Simétrica
Tempo Médio=10,2
Frequency
0
1
2
3
4
5
6
7
5
10
15
Tempo de reação

- Tempo médio de reação do teste sorológico em homens e mulheres.

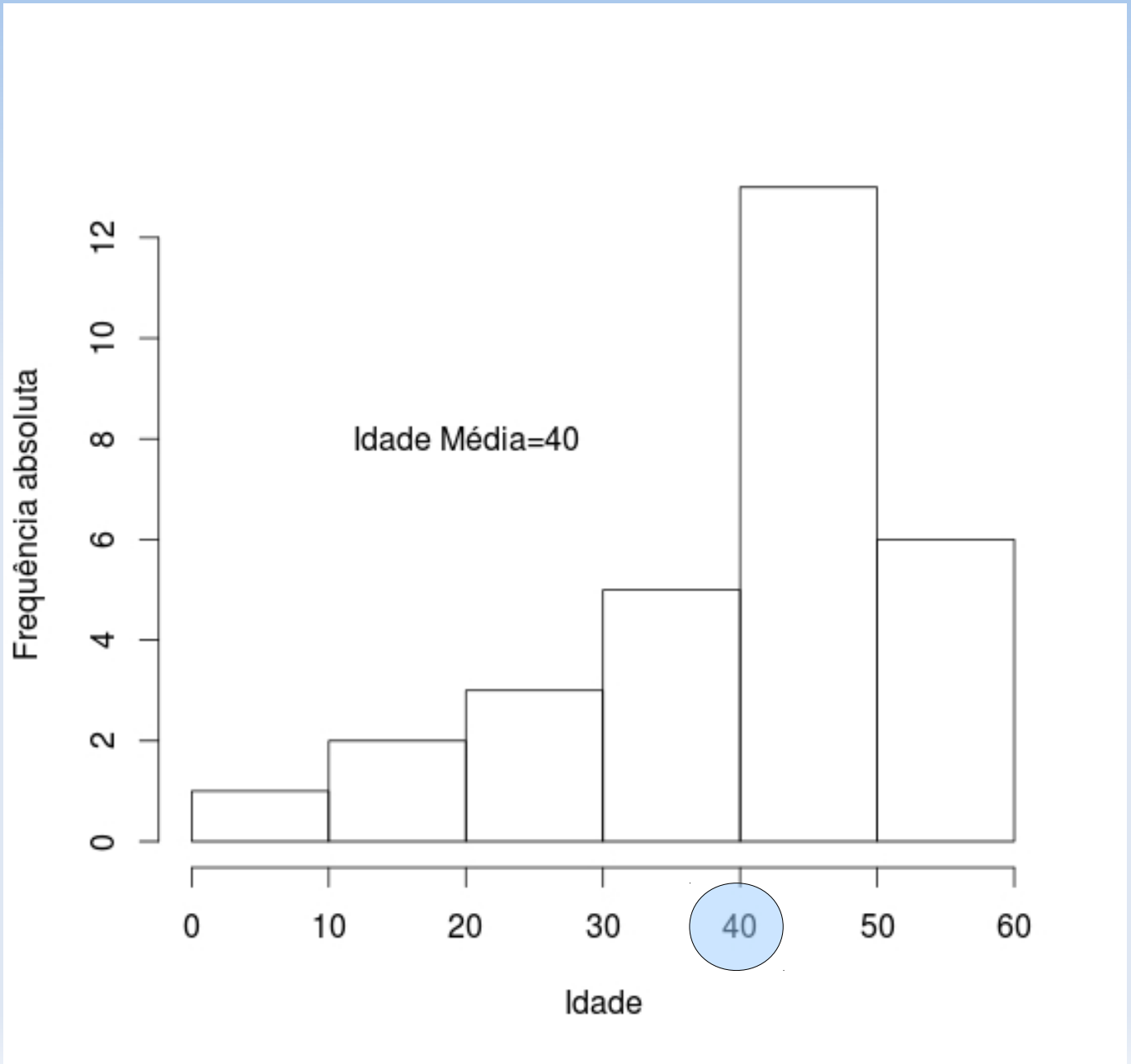
	Feminino	Masculino
	8,7	15,5
	11,9	2,8
	5	9,7
	11,1	13
	10,7	7,1
	11,7	5,7
	13,3	6,3
	8,8	15,1
	8,3	16,9
	7,8	9,1
	13,1	13,5
	15,4	11,2
	10,8	3,6
	9,8	12,8
	7,2	
	10,6	
Soma	164,2	142,3
n	16	14
Média	10,26	10,16

Tempo de reação segundo categorias de reação e de sexo

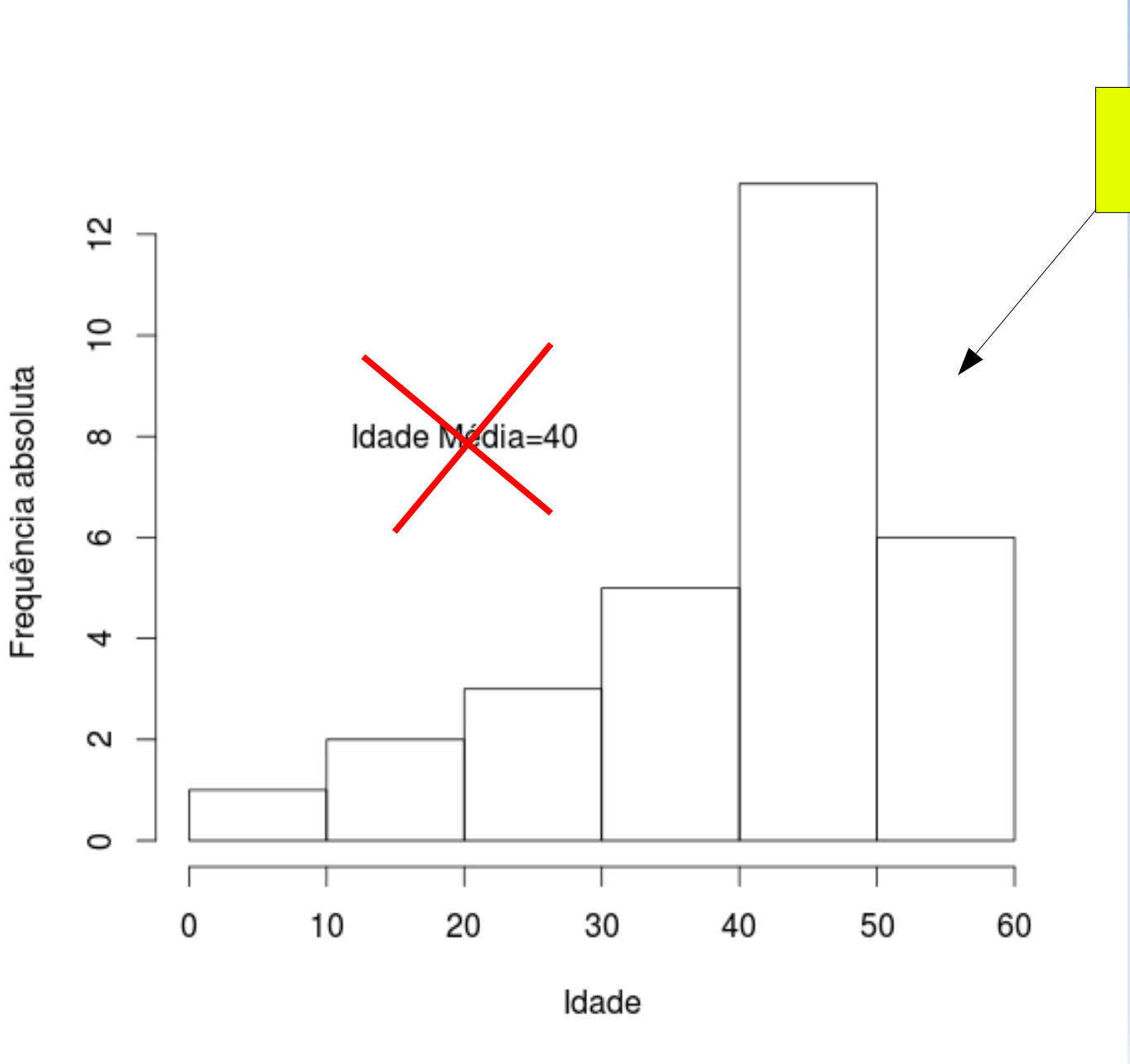
	Sexo		
Reação	feminino	masculino	Média
positiva	9,26	10,74	9,87
negativa	11,04	9,84	10,44
Média	10,26	10,16	

Interpretação ?

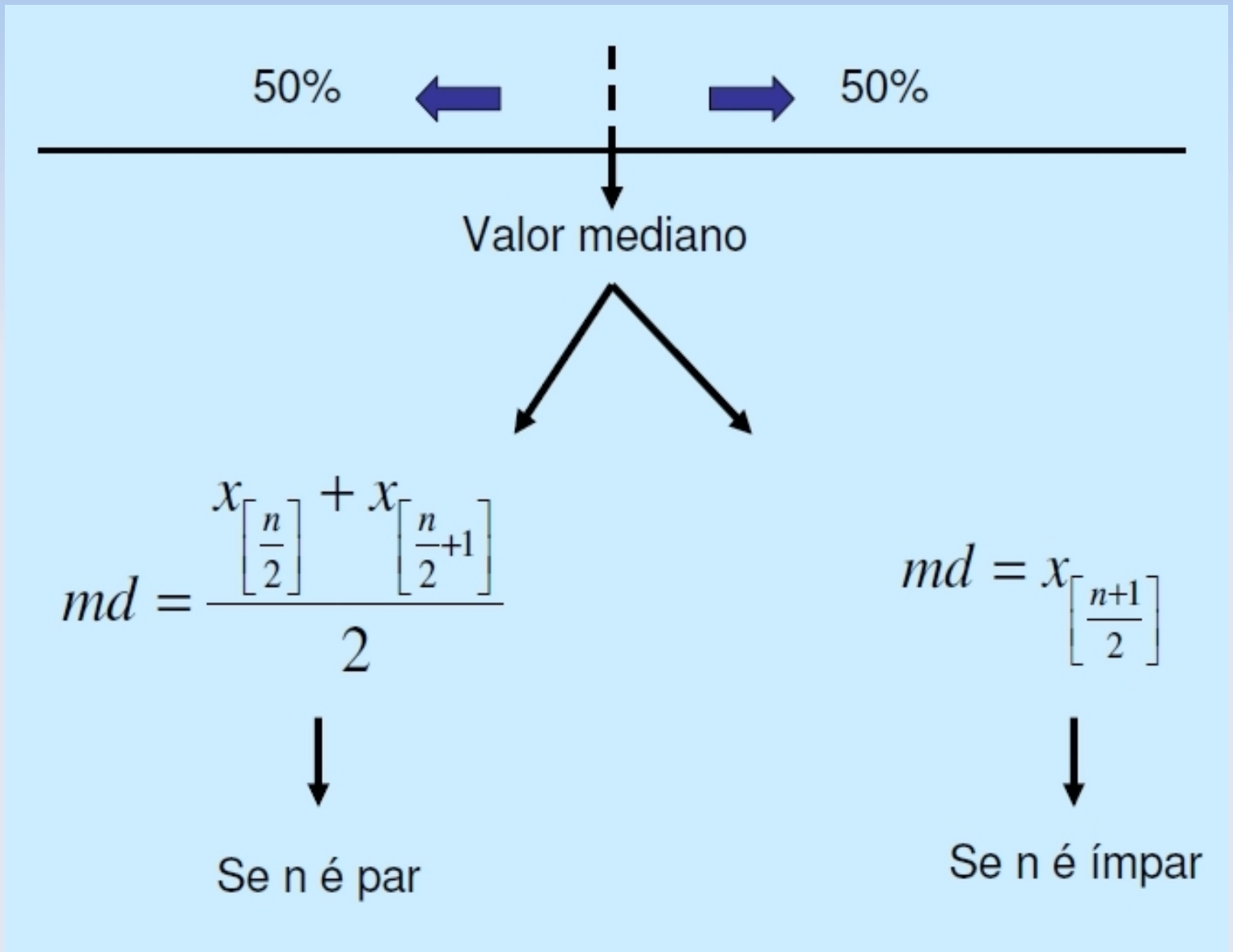
O problema da distorção



O problema da distorção

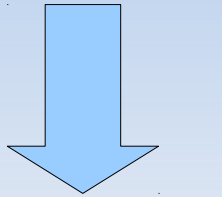


Mediana


50%
50%
Valor mediano
$md = \frac{x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}}{2}$
$md = x_{\left[\frac{n+1}{2}\right]}$
Se n é par
Se n é ímpar

Exemplo: idade mediana

8 11 19 21 22 28 34 34 35 35 38 42 42 42 44 44 45 45 46 46 46 48 50 50 52 52 56 56 58 58



Md=44

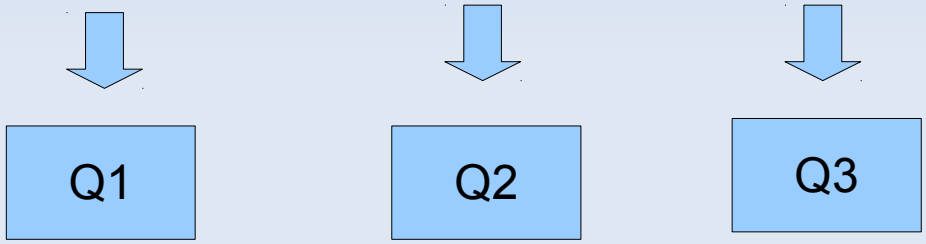
Interpretação ?

Quartis e amplitude inter-quartis

- Uma outra forma de sumarizar dados é em termos dos quantis.
- São particularmente úteis para dados não simétricos.
- **quartis inferior e superior** (Q1 e Q3) são os valores abaixo dos quais estão um quarto e três quartos dos dados.
- **mediana** (Q2) é o valor que divide os dados ordenados ao meio
- Estes três valores são usados para resumir os dados junto com o **mínimo** e o **máximo**.

- Eles são obtidos ordenando os dados do menor para o maior, e conta-se o número apropriado de observações:

(n+1)/4, (n+1)/2 e 3(n+1)/4



- Para um número par de observações, a mediana é a média dos valores do meio (e analogamente para os quartis inferior e superior).
- A medida de dispersão é a amplitude inter-quartis: **IQR=Q3-Q1**

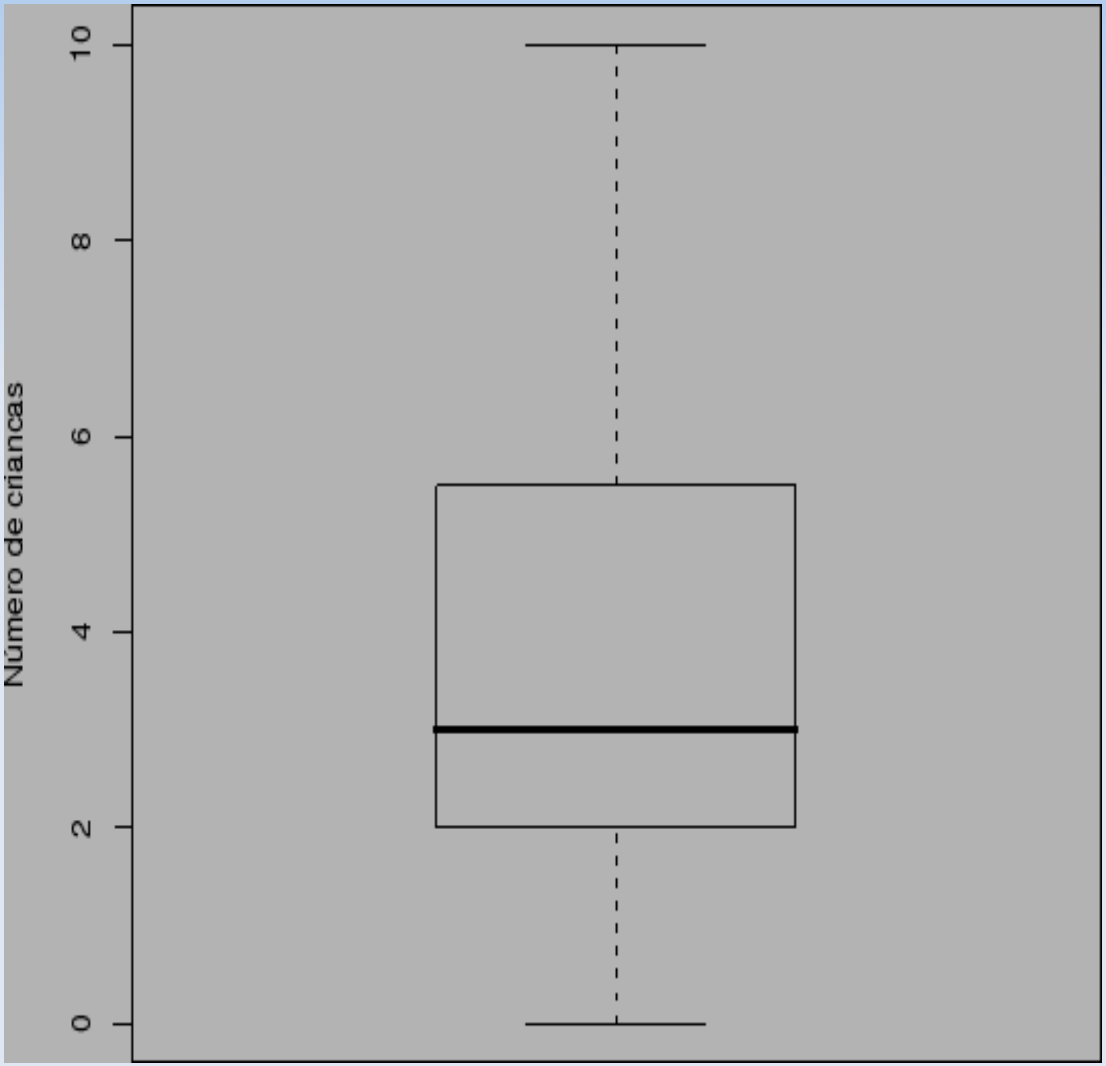
Exemplo

- O número de crianças em 19 famílias foi

0, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 7, 8, 10

- A mediana é o (19+1)/2 =10o. valor: **Q2=3** crianças.
- O quartil inferior é o (19+1)/4=5o. valor e o quartil superior é o 3(19+1)/4=15o.: **Q1=2 e Q3=6** crianças
- Amplitude inter-quartis é de 4 crianças.
- Note que 50% dos dados estão entre Q1 e Q3.

Box Plot



Box-plots são representações diagramáticas dos cinco números sumários: (mínimo, quartil inferior, mediana, quartil superior, máximo).

Medidas de variabilidade

- Amplitude total

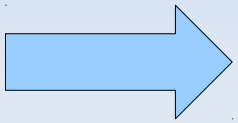
  A=Máx-Min

  - Exemplo: Amplitude das idades = 58-8 = 50

É uma boa medida de variabilidade?

Variância

- Considere os conjuntos:
  - A={3,4,5,6,7}
  - B={1,3,5,7,9}
  - C={5,5,5,5,5}
  - D={3,5,5,7}



Média = 5

- O conjunto C não apresenta variação. Uma medida óbvia seria …
- Os conjuntos A, B e D têm variação. Como mensurá-las?

A idéia é "medir" a dispersão dos dados em relação à média

Desvios

QUADRO DOS DESVIOS

Desvios	A	B	C	D
	-2	-4	0	-2
	-1	-2	0	0
	0	0	0	0
	1	2	0	2
	2	4	0	
Soma				

A idéia é "medir" a dispersão dos dados em relação à média

Desvios

QUADRO DOS DESVIOS

Desvios	A	B	C	D
	-2	-4	0	-2
	-1	-2	0	0
	0	0	0	0
	1	2	0	2
	2	4	0	
Soma	0	0	0	0

Quadro dos desvios quadráticos

Desvios Quadráticos	A	B	C	D
	4	16	0	4
	1	4	0	0
	0	0	0	0
	1	4	0	4
	4	16	0	
Soma	10	40	0	8
Desvio Quadrático Médio	**2**	**8**	**0**	**2**

Quadro dos desvios quadráticos

Desvios Quadráticos	A	B	C	D
	4	16	0	4
	1	4	0	0
	0	0	0	0
	1	4	0	4
	4	16	0	
Soma	10	40	0	8
Desvio Quadrático Médio	**2**	**8**	**0**	**2**

VARIÂNCIA

Definição de variância

- N: total populacional

$$\sigma^2 = \frac{\sum \left(x - \bar{x}\right)^2}{N}$$

- n: total amostral

$$s^2 = \frac{\sum \left(x - \bar{x}\right)^2}{n - 1}$$

Exemplo

- Considere os conjuntos:
  - A={3,4,5,6,7} $s^2=2,5$
  - B={1,3,5,7,9} $s^2=10$
  - C={5,5,5,5,5} $s^2=0$
  - D={3,5,5,7} $s^2=2,7$

Exemplo: teste sorológico

negativa	positiva
15,5	8,7
2,8	11,9
13,0	5,0
7,1	9,7
11,1	6,3
5,7	15,1
10,7	8,8
11,7	9,1
13,3	7,8
8,3	13,5
16,9	15,4
13,1	7,2
10,8	
11,2	
3,6	
9,8	
12,8	
10,6	

Soma	188,00	118,50
Soma quad	2204,86	1296,43
n	18	12

Comparar os tempos de reação em ensaios com resultados positivos e negativos

	negativa	positiva
média	**10,44**	**9,88**
variância	**14,19**	**11,48**

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \quad \Rightarrow \quad s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

Coeficiente de variação

$$C.V = \frac{s}{\overline{x}} 100 \quad (\%)$$

Medida de dispersão relativa (pura)

Ex: Comparar a variabilidade das idades com os tempos de reação

	idade	Tempos de reação
média	40,23	10,22
desvio-padrão	13,36	3,57
CV		

Coeficiente de variação

$$C.V = \frac{s}{\overline{x}} 100 \quad (\%)$$

Medida de dispersão relativa (pura)

Ex: Comparar a variabilidade das idades com os tempos de reação

	idade	Tempos de reação
média	40,23	10,22
desvio-padrão	13,36	3,57
CV	33	35

Escore padronizado

- Ao contrário do CV, é útil para comparação de resultados indivíduais.
- Por exemplo compare:

Nota	Média	Desempenho
7	5	
8	9	

- Além da comparação da nota individual com a média da turma, é importante avaliar se a variabilidade foi grande ou não.
- Por exemplo:

Nota	Média	Desvio-padrão	Desempenho
7	5	2	
7	5	4	

Escore padronizado

$$Z = \frac{x - \bar{x}}{s}$$

Nota	Média	Desvio-padrão	Escore Padronizado
7	5	2	1
7	5	4	0,5

Interpretação?