SS714 - Bioestatística

Silvia Shimakura

silvia.shimakura@ufpr.br

Página da disciplina:

http://www.leg.ufpr.br/doku.php/disciplinas:ss714

ESTATÍSTICA DESCRITIVA

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

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

DADOS (OU VARIÁVEIS)

- Quantificação ou categorização do fenômeno de interesse

Inquérito epidemiológico:

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

Tipos de Dados

- Facilita o tratamento estatístico classificar dados em: **Qualitativos e Quantitativos**

- **Qualitativos**
  - **Nominais:** Emprego, Estado civil
  - **Ordinais:** Grau de instrução, Faixa de renda

- **Quantitativos**
  - **Discretas:** Tamanho da família, Renda
  - **Contínuas:** Idade, Renda

Banco de dados

- Uma linha para cada indivíduo
- Uma coluna para cada variável observada
- Para variáveis qualitativas:
  - Criar códigos para cada categoria
- Para variáveis contínuas:
  - Entrar com os dados originais e não os codificados para classes de interesse (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: Tentativas de suicídio

- Estudo retrospectivo (Fernandes et al., 1995)
- Tentativas de suicídio por intoxicação aguda registradas no Centro de Assistência Toxicológica do Hosp. de Base de São Paulo.
- Período de 01/92 a 02/93: 302 casos

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

- **Dicionário das variáveis:**
- **Sexo:** 0=Masculino e 1=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

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.

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

Distribuição de tentativas de suicídio segundo faixa etária

Idade (anos)	Frequência		
	Absoluta	Relativa (%)	Acumulada (%)
10⊢20	59	19,54	19,54
20⊢30	115	38,08	57,62
30⊢40	61	20,20	77,82
40⊢50	35	11,59	89,41
50⊢60	21	6,95	96,36
60⊢70	9	2,98	99,34
70⊢80	2	0,66	100
Total	302	100	

Etapas para construção de tabelas de frequências para dados agrupados

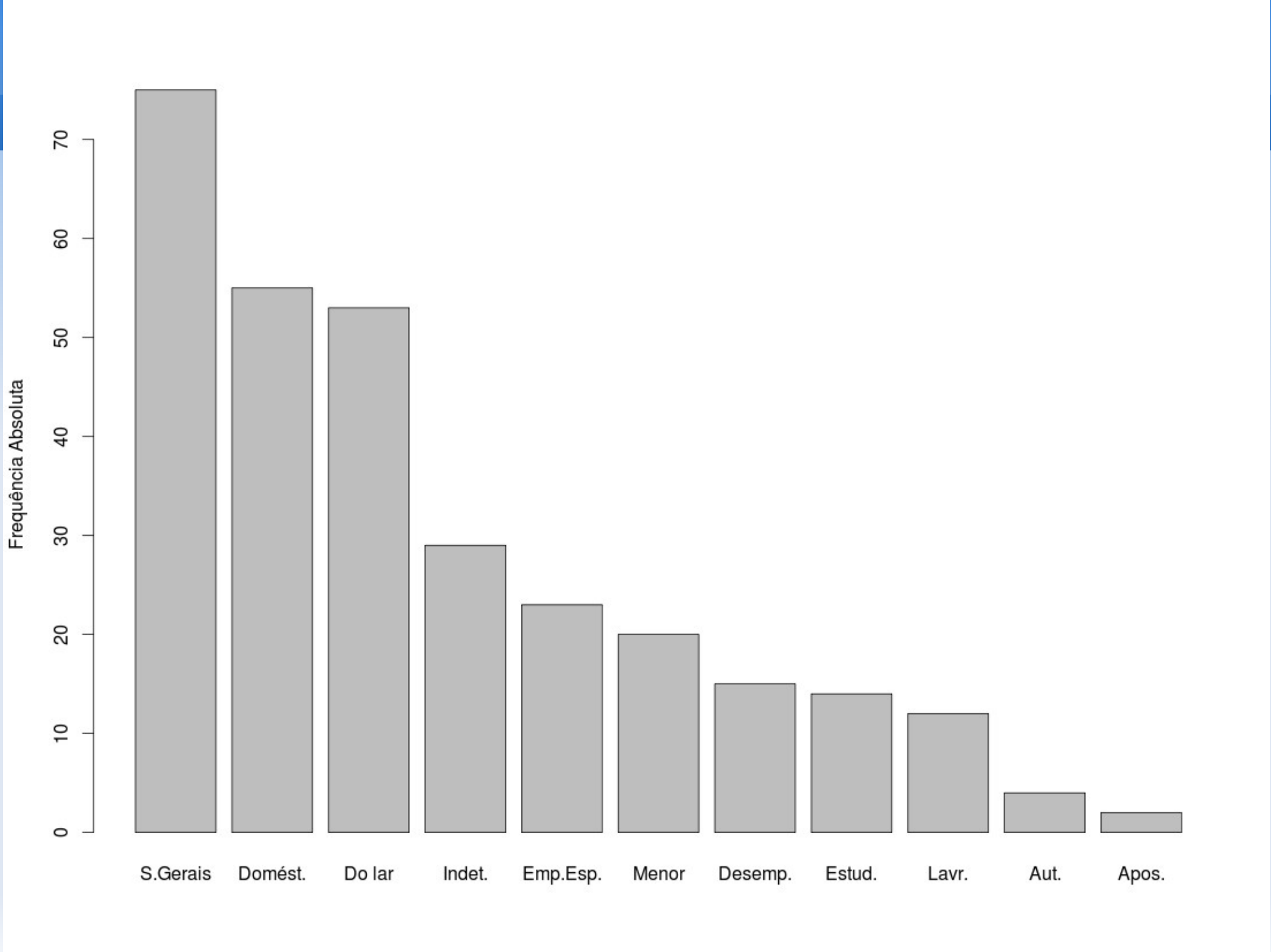
1. Encontrar o menor e o maior valores (mínimo e máximo) do conjunto de dados
2. Escolher número de classes (de igual amplitude), que englobem todos os dados sem superposição de intervalos.
3. Contar o número de elementos em cada classe (este número é a frequência absoluta)
4. Calcular a frequência relativa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

GRÁFICOS

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

Representação gráfica para variáveis categóricas

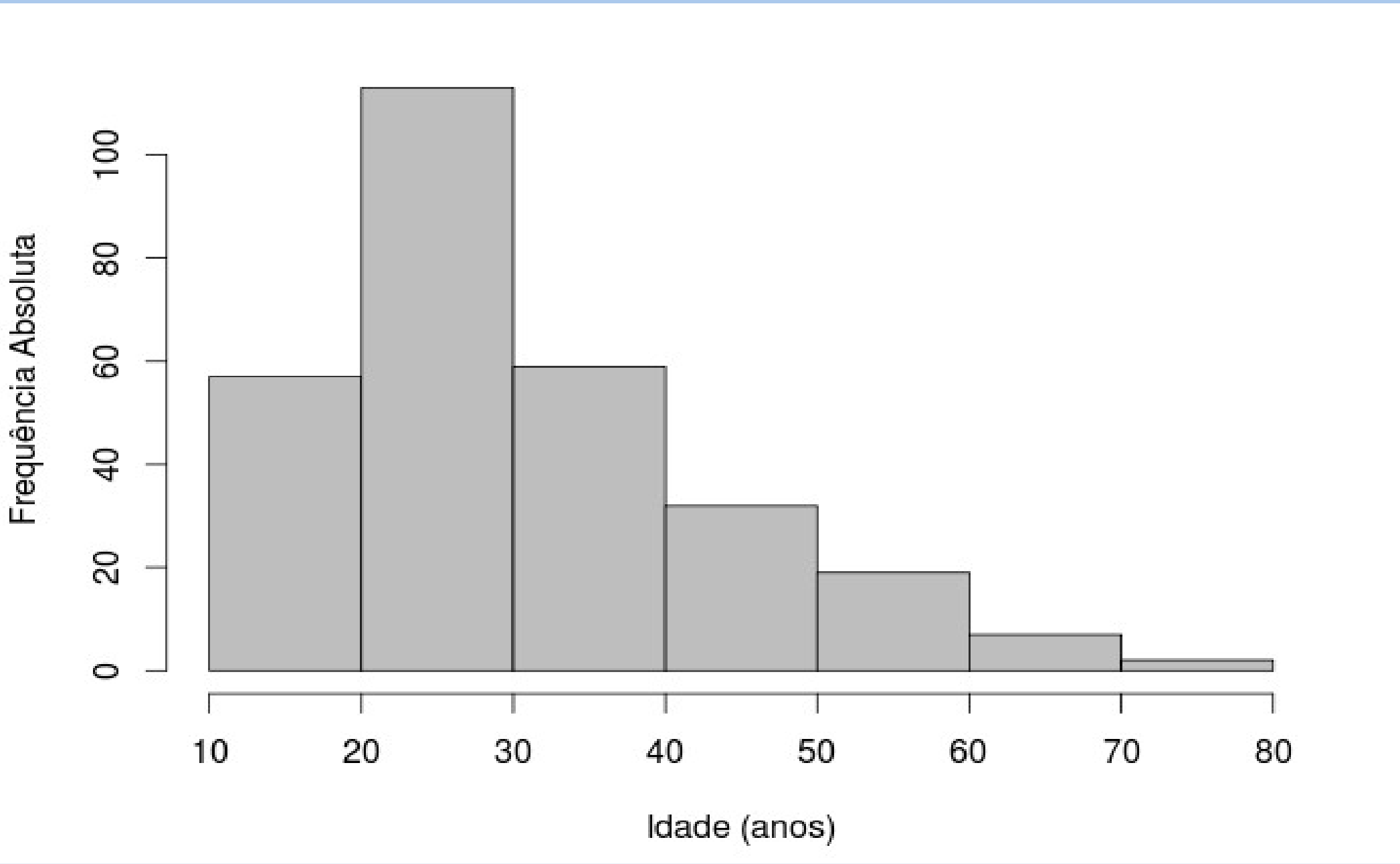
- Diagrama de barras
- Exemplo 3.5: Distribuição de profissões entre pacientes potencialmente suicídas (cont.)


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.

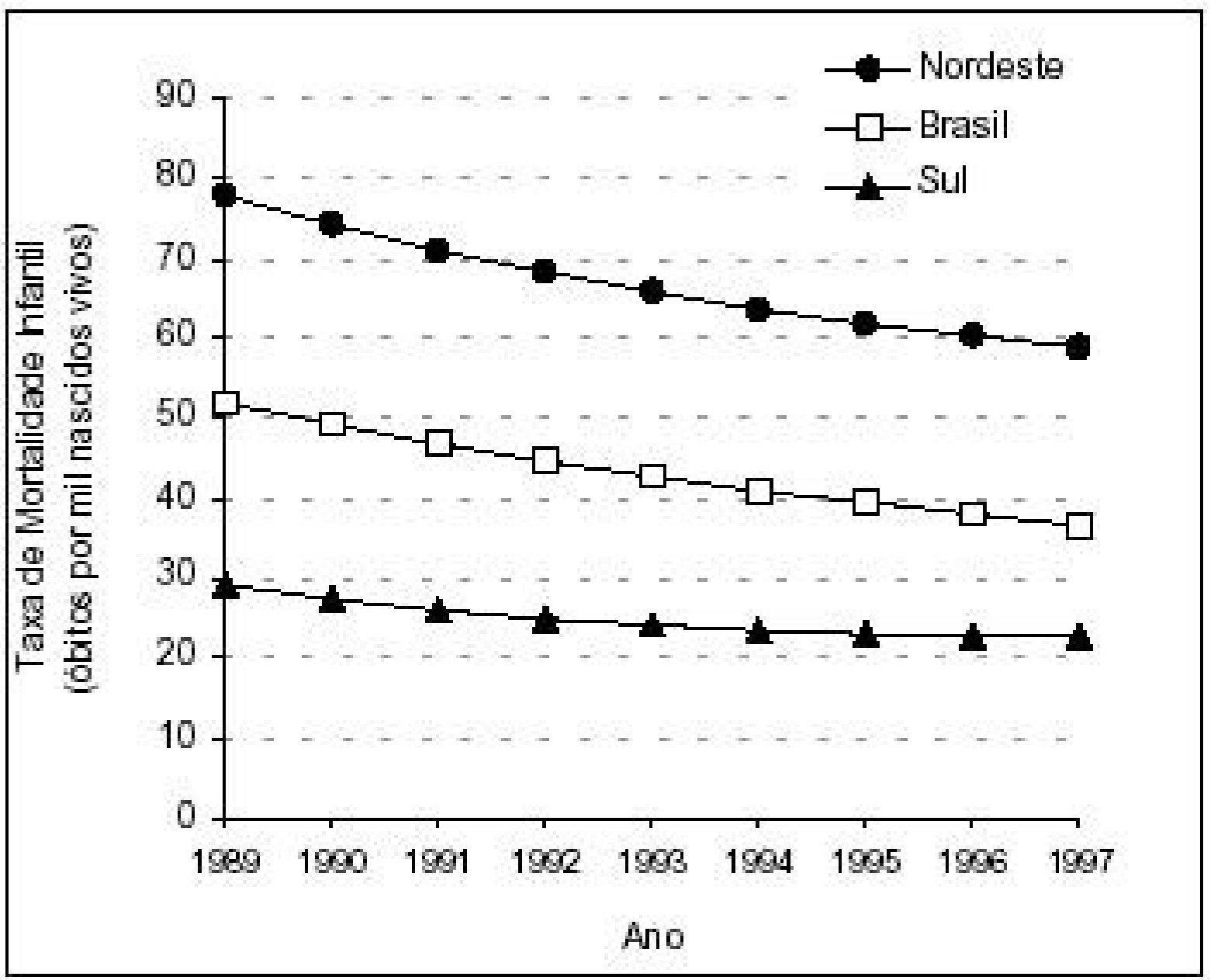
Exemplo 3.5: Distribuição das tentativas de suicídio segundo faixa etária



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



RESUMOS NUMÉRICOS

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

Dados Qualitativos

- Para resumir dados qualitativos numericamente usamos contagens, proporções, taxas
- Exemplos:
  - Se 70 de 140 estudantes de medicina são mulheres, podemos dizer que a proporção de mulheres é de 0,5 ou em termos percentuais que 50% são mulheres.
  - Se numa amostra de 5000 pessoas, 7 são portadores de uma doença podemos expressar este achado como uma proporção (0,0014) ou percentual (0,14%), ou taxa (1,4 por mil).

Exemplo: Tentativas de suicídio

- 302 casos: 27% do total de atendimentos no período
- 67% das tentativas de suicídio do sexo feminino

Dados Quantitativos

- Para resumir numericamente dados quantitativos escolhemos medidas de:
  - **Locação (Tendência Central)**

    Valor ao redor do qual as observações tendem a se agrupar
  - **Dispersão (Variabilidade)**

    As observações estão próximas do centro ou estão dispersas num amplo intervalo de valores?
- Existem três medidas principais de locação e dispersão:

Locação	Dispersão
Média	Desvio-padrão
Mediana	AIQ
Moda	Proporção

Moda e Proporção

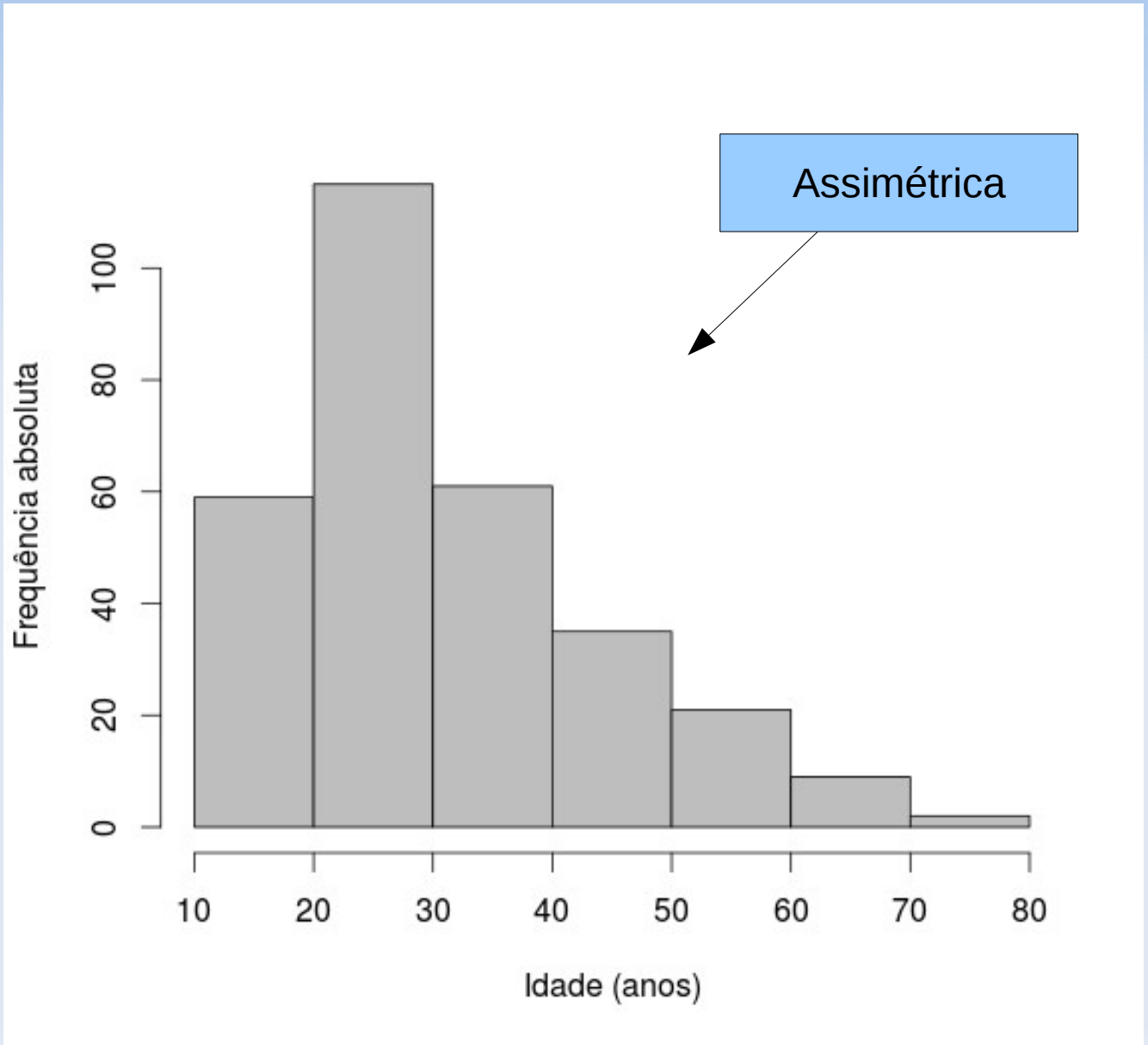
- **Moda:** Valor mais que ocorre com mais frequência
- **Dispersão:** Proporção dos dados iguais à moda

Distribuição de tentativas de suicídio segundo faixa etária

Classe modal →

Idade (anos)	Frequência		
	Absoluta	Relativa (%)	Acumulada (%)
10⊢20	59	19,54	19,54
20⊢30	115	**38,08**	57,62
30⊢40	61	20,20	77,82
40⊢50	35	11,59	89,41
50⊢60	21	6,95	96,36
60⊢70	9	2,98	99,34
70⊢80	2	0,66	100
Total	302	100	

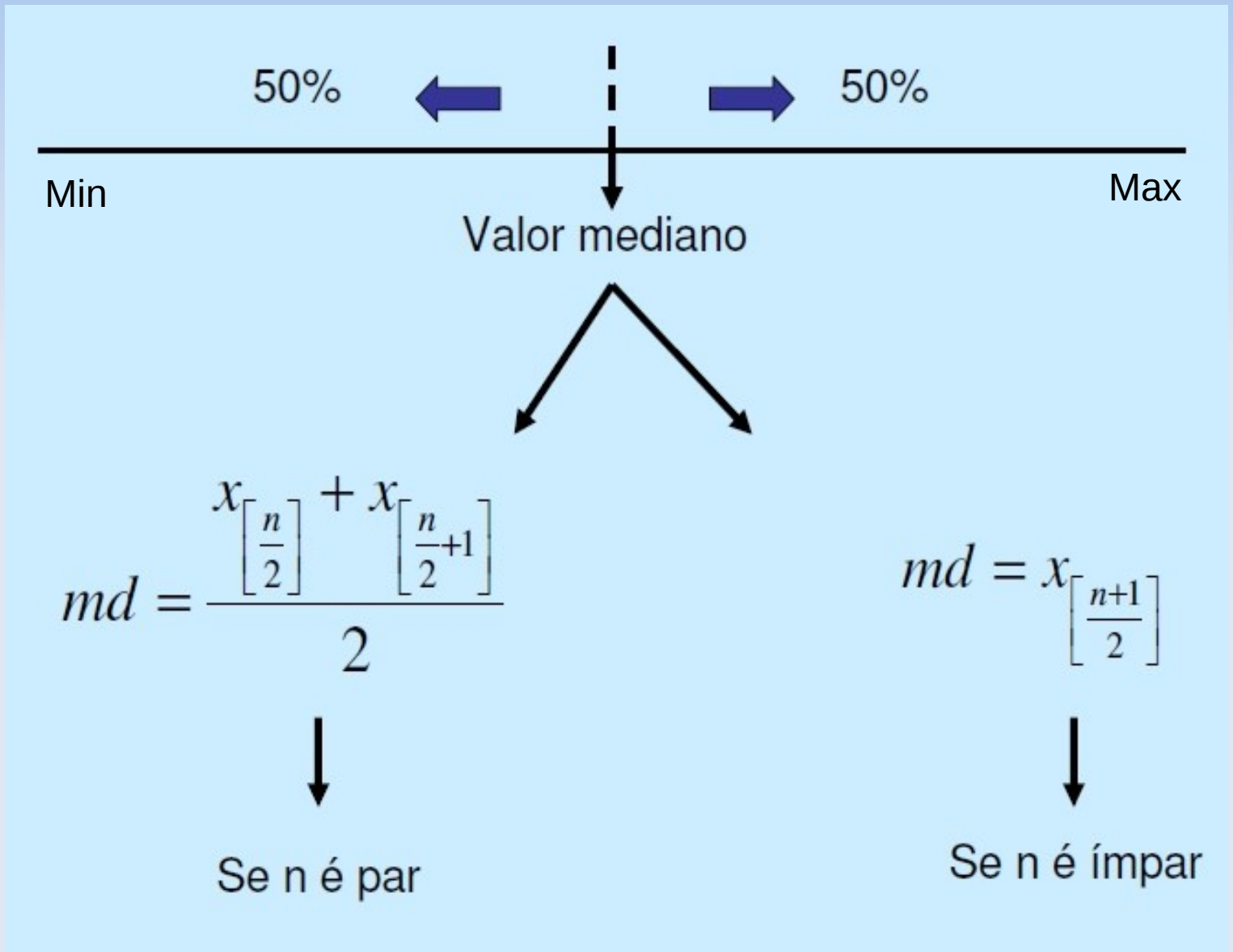
Problema da distorção



Mediana e AIQ

- **Quartis ou Percentis:** especialmente úteis para dados não simétricos
- **Mediana (ou Percentil 50):** valor que divide os dados ordenados ao meio, ou seja, ½ dados tem valores maiores do que a mediana, ½ dados tem valores menores do que a mediana.
- **Quartis inferior e superior (Q1 e Q3):** valores baixo dos quais caem ¼ e ¾ dos dados.
- **5 números sumários (MQMQM):** Min, Q1, Mediana, Q3, Max
- **Amplitude Inter-Quartis:** AIQ=Q3-Q1

Mediana


50%
50%
Min
Max
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

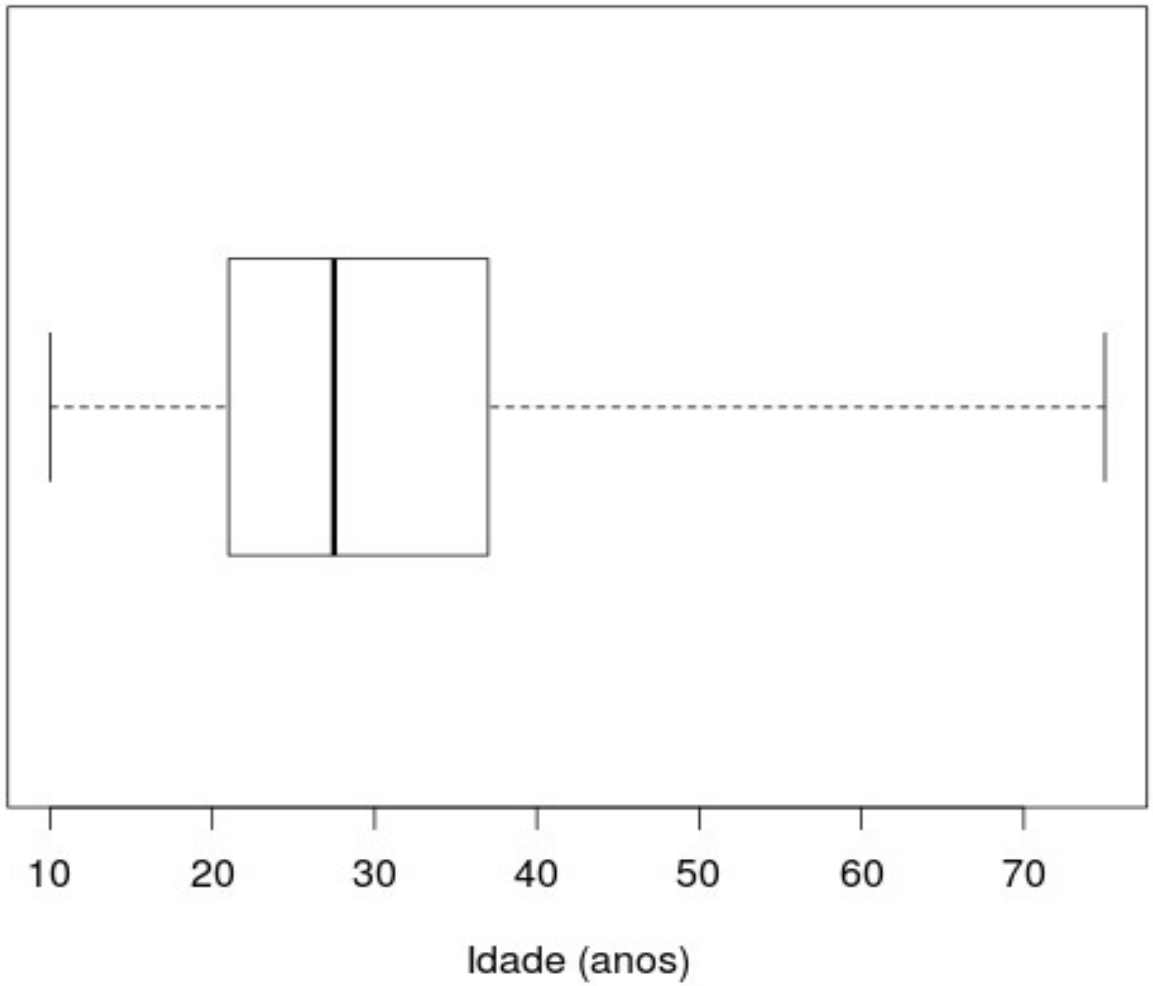
Usando o R

```
> summary(idade)
```

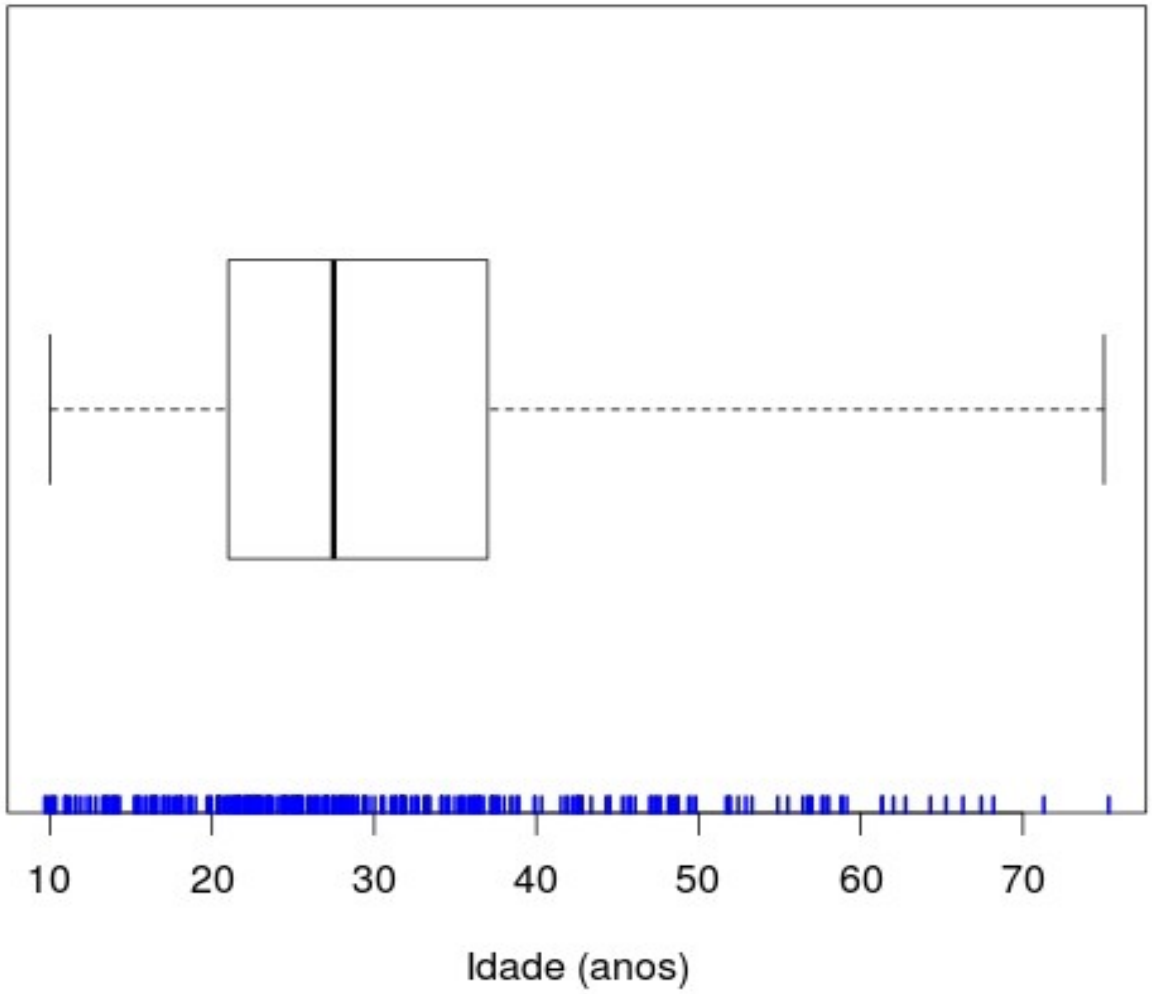
Min.	1st Qu.	Median	3rd Qu.	Max.
10.00	21.25	27.50	37.00	75.00

```
> boxplot(idade, range=0,xlab='Idade (anos)',horizontal=TRUE)
> rug(jitter(idade,amount=0.5), col='blue')
```

Boxplot das idades



Boxplot das idades

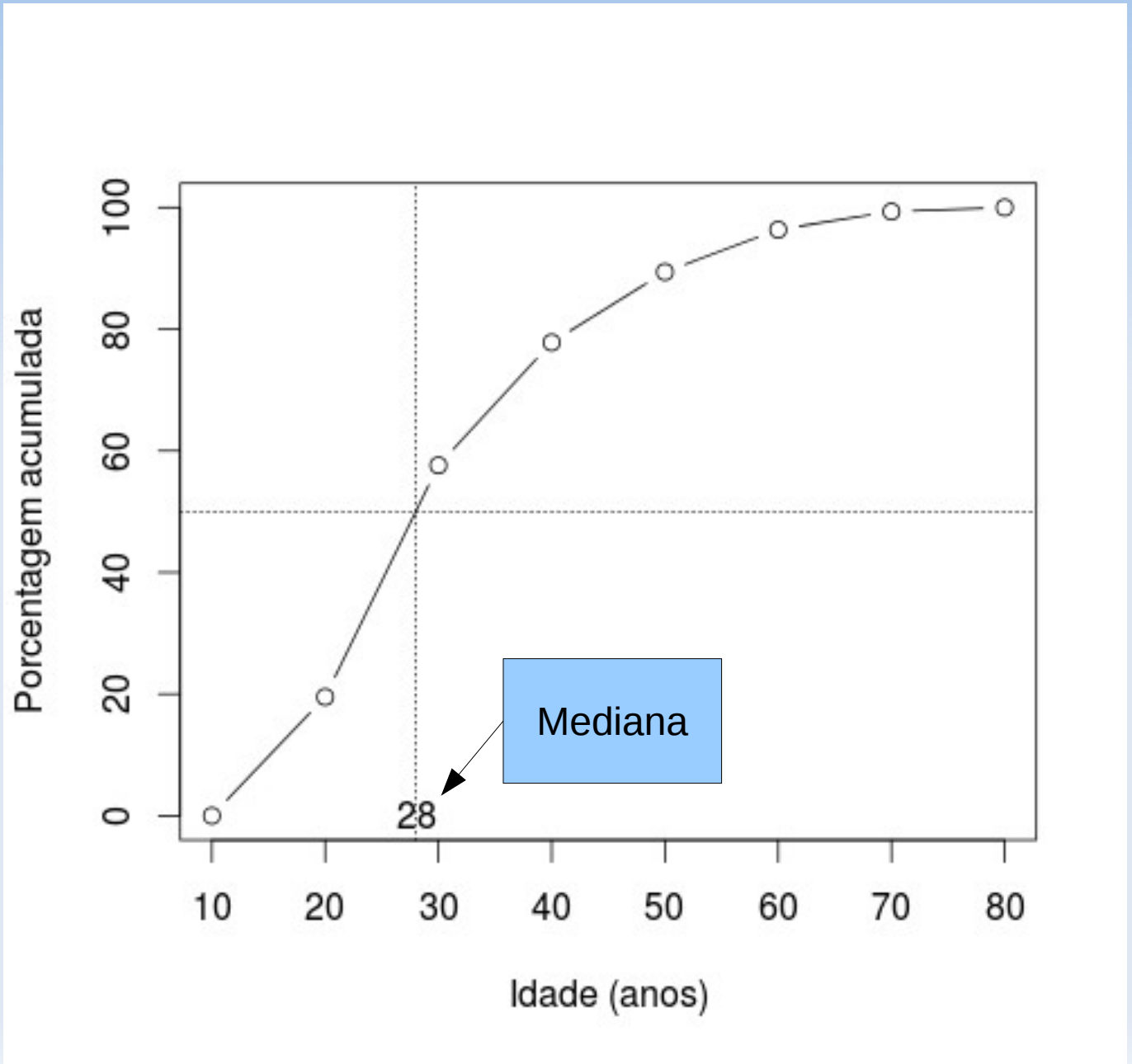


Ogiva

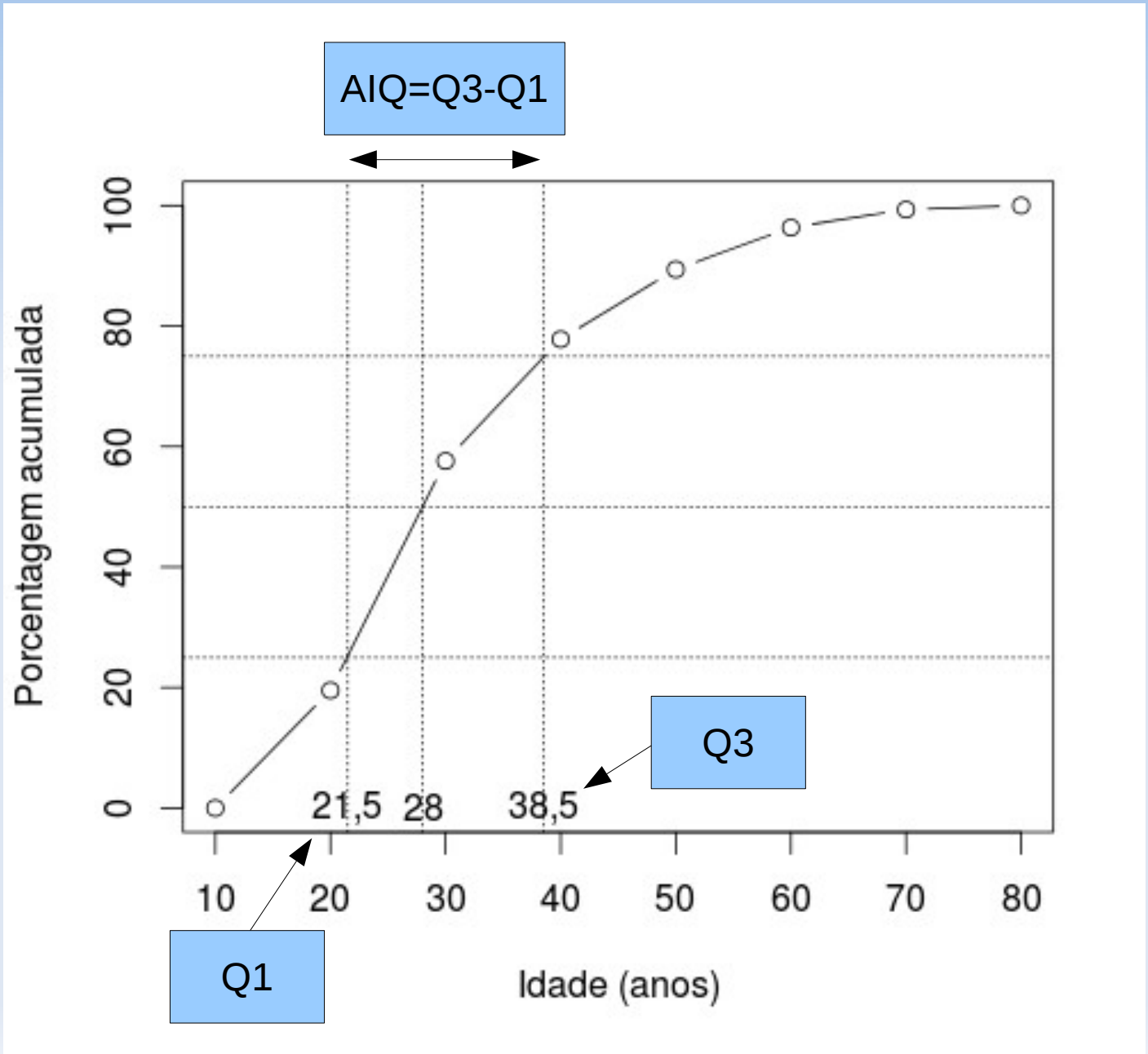
- Gráfico de percentuais acumulados
- Através da ogiva podemos estimar qualquer percentil da distribuição.

- **Exemplo:** Estimar a idade abaixo da qual encontram-se 50% dos indivíduos.

Ogiva das idades



Ogiva das idades



Exemplo: Teor de gordura fecal

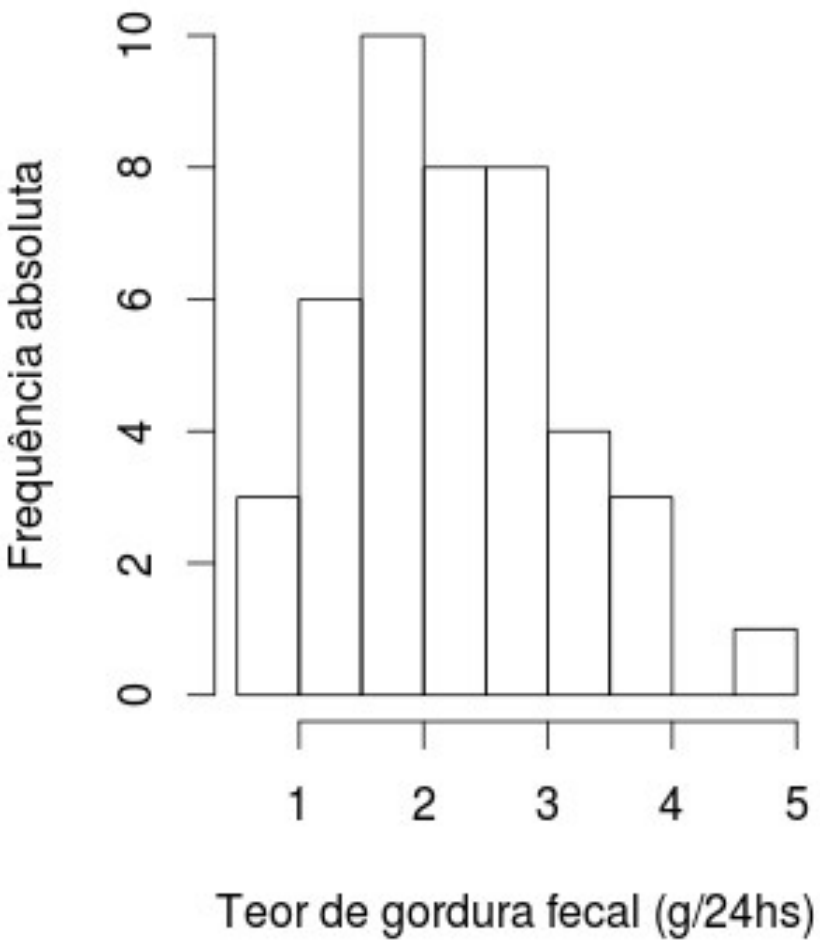
- **Dosagem de gordura:** útil no diagnóstico e acompanhamento da síndrome de má absorção - quando se tem a síndrome tem-se um aumento do teor de gordura fecal.
- Até 1984 não existia um padrão de referência para crianças brasileiras.
- Prof. Francisco Penna (titular de pediatria da UFMG) examinou 43 crianças sadias

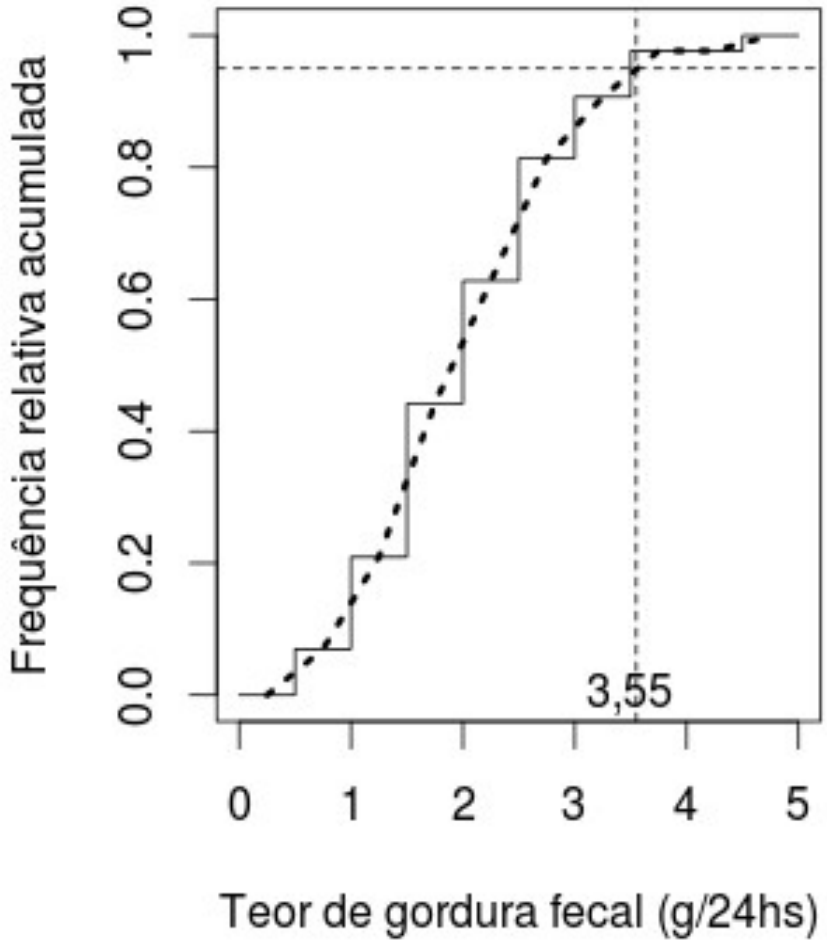
Tabela: Teor de gordura fecal (g/24 hs)

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 variabilidade dos resultados!
- Podemos definir um padrão de referência usando a ogiva.

Exemplo: Teor de gordura fecal em crianças sadias





Média e desvio-padrão

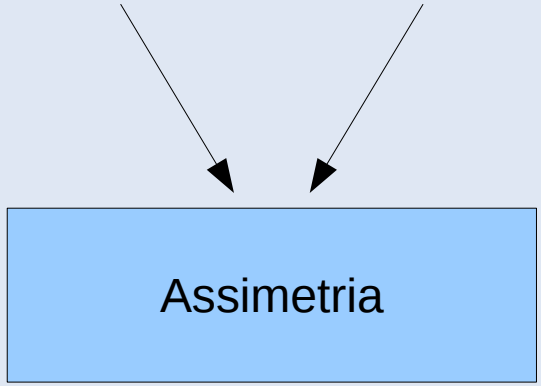
- Usada para resumir dados quantitativos simétricos

- **Média:**

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

Exemplo: Tentativas de suicídio (cont.)

```
> summary(idade)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   21.25   27.50   30.61   37.00   75.00
```



Medidas de variabilidade

- Amplitude total

$$A = Máx - Min$$

- Exemplo: Amplitude das idades = 75-10 = 65

É uma boa medida de variabilidade?

Medidas de variabilidade

- Amplitude total

$$A = Máx - Min$$

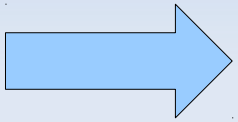
- Exemplo: Amplitude das idades = 75-10 = 65

É uma boa medida de variabilidade?

Não utiliza todas as observações.

Medidas de variabilidade

- 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 …
- Como medir variação nos conjuntos A, B e D?

Desvio médio

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				

Desvio quadrático médio

	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				

Desvio quadrático médio

	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

	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

Desvio quadrático médio

	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	
	10	40	0	8
Desvio quadrático médio	2	8	0	2

VARIÂNCIA

Definição de variância

- N: total populacional

Variância populacional →

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

- n: total amostral

Variância amostral →

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

Exemplo

- Considerando que A, B, C e D são amostras:
  - 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} $s^2=2,7$

Desvio-padrão

- A variância é uma medida de dispersão obtida numa escala quadrática.
- Para que a dispersão tenha a mesma unidade de medida dos dados originais calculamos a raiz quadrada da variância.

Desvio-padrão=√variância

- σ=desvio-padrão populacional
- s=desvio-padrão amostral

Exemplo

- Considerando que A, B, C e D são amostras:
  - A={3,4,5,6,7} $s^2=2,5$ $s=\sqrt{2,5}=1,58$
  - B={1,3,5,7,9} $s^2=10$ $s=\sqrt{10}=3,16$
  - C={5,5,5,5,5} $s^2=0$ $s=\sqrt{0}=0$
  - D={3,5,5,7} $s^2=2,7$ $s=\sqrt{2,7}=1,64$

Exemplo: Teste sorológico

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
...	...	...	...	...	...
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

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**

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

Exemplo: Teste sorológico

Ex: Os pacientes são mais parecidos entre sí nas idades ou nos tempos de reação?

	Idade	Tempo de reação
média	40,23	10,22
desvio-padrão	13,36	3,57

Coeficiente de Variação

Ex: Os pacientes são mais parecidos entre sí nas idades ou nos tempos de reação?

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

$$CV = \frac{s}{\bar{x}} 100 \quad (\%)$$

Medida de dispersão relativa (pura)

Coeficiente de Variação

Ex: Os pacientes são mais parecidos entre sí nas idades ou nos tempos de reação?

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

$$CV = \frac{s}{\bar{x}} 100 \quad (\%)$$

Medida de dispersão relativa (pura)

Escore padronizado

- Ao contrário do CV, é útil para medir resultado indivídual.
- Por exemplo compare:

Nota	Média	Desempenho
7	5	
8	9	

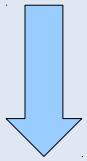
- Além de comparar a nota individual com a média da turma, é importante avaliar se a variabilidade foi grande ou não.
- Por exemplo:

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?

Exercício

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

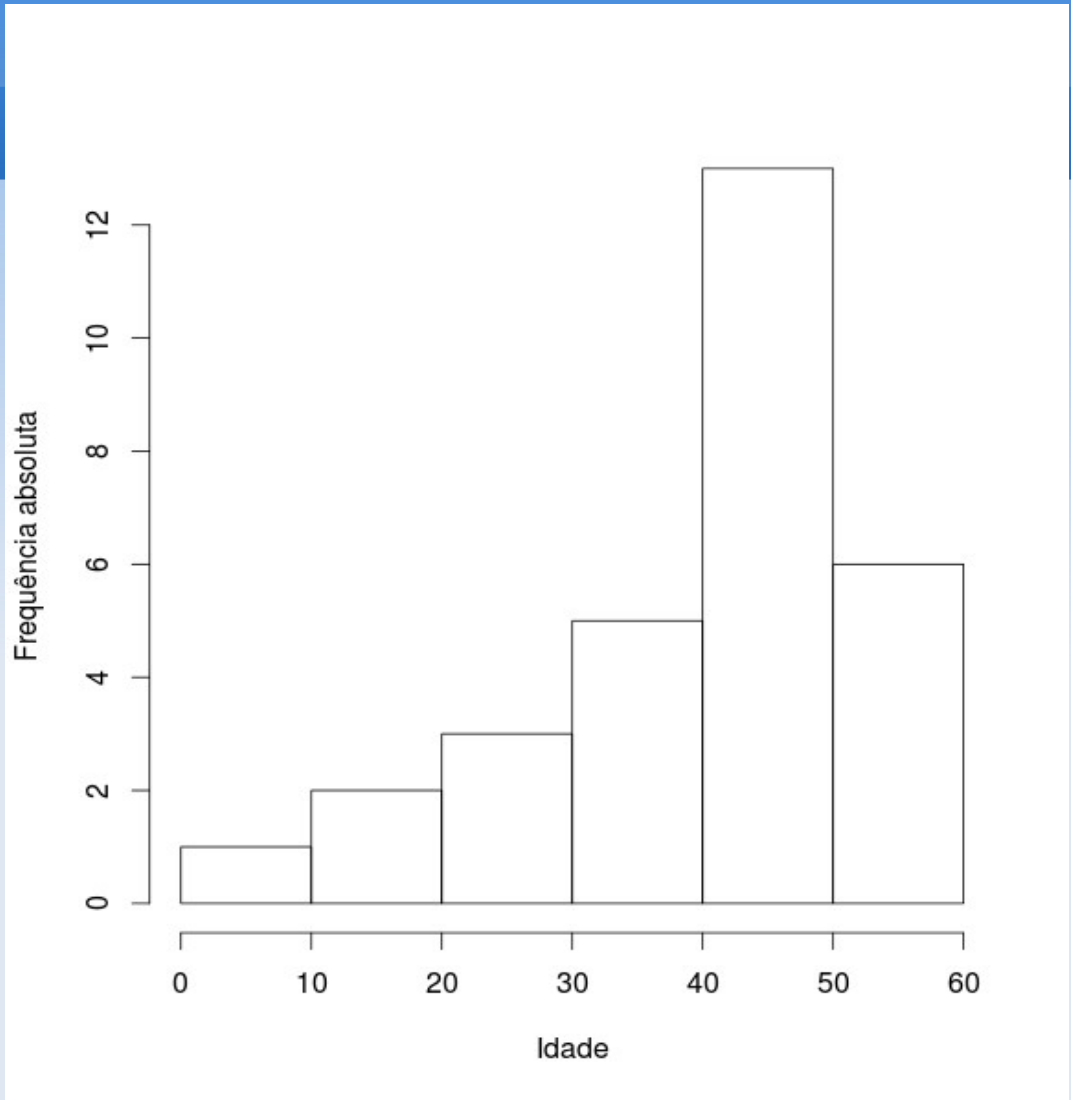
Característica ou valor que ocorre com maior freqüência.

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

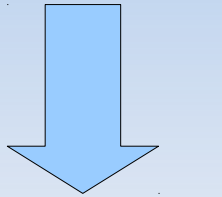
Com dados quantitativos de natureza contínua, em geral, basta identificar a classe modal.



Classe modal

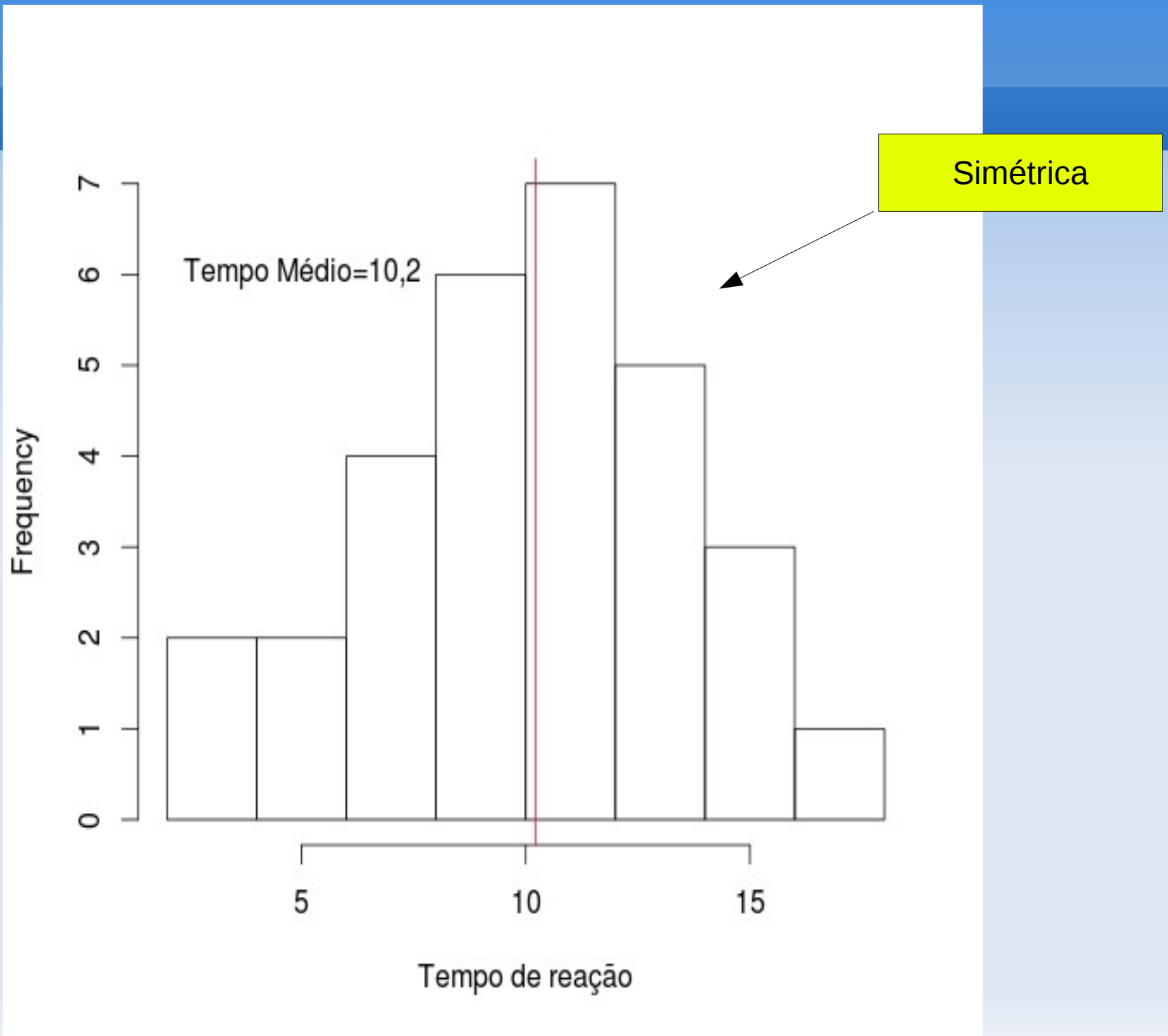
Exemplo: idade mediana

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 ?


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.
- Extra: Analisar o tempo de reação segundo as categorias positivo ou negativo.

	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

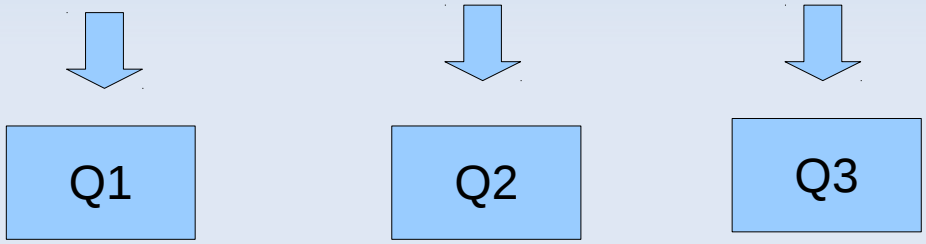
	Sexo		
Reação	feminino	masculino	Média
positiva	9,26	10,74	9,87
negativa	11,04	9,84	10,44
Média	10,26	10,16	

Interpretação ?

Quartis para dados não agrupados

- 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**

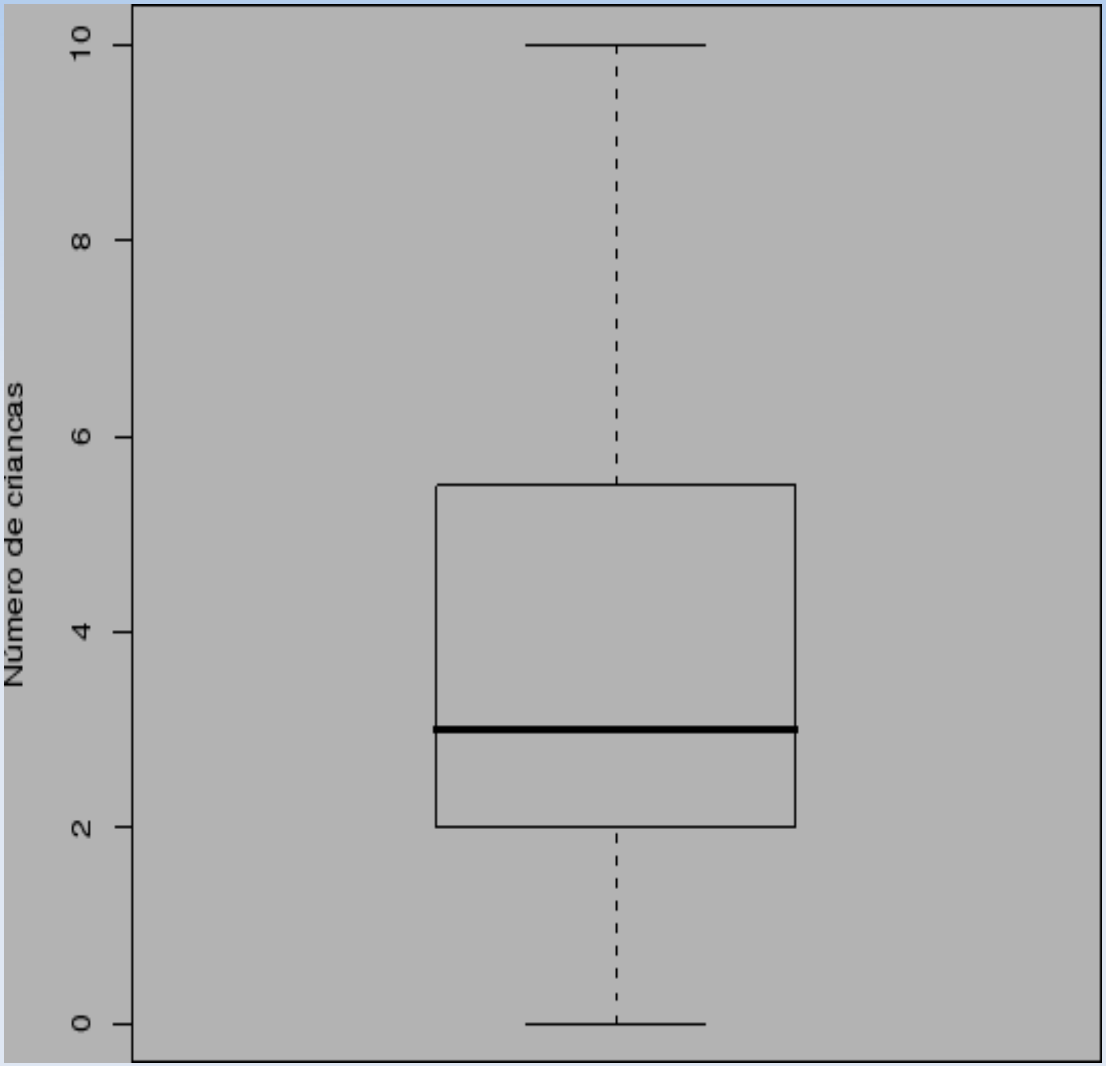
Exemplo

- 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 5o. valor e o quartil superior é o 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).