

SCOL7000 - Bioestatística

Silvia Shimakura

`silvia.shimakura@ufpr.br`

Página da disciplina:

<http://www.leg.ufpr.br/doku.php/disciplinas:scol7000>

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 (você pode querer mudar as classes durante a análise)
- Para **dados omissos**: deixar os campos em branco ou usar código que facilmente identifique esse tipo de dado (Ex: 999 para pressão arterial)

Exemplo: Peso de recém nascidos (birthwt.dat)

- Dados de 189 nascimentos num hospital dos EUA
- Principal interesse era em recém nascidos com baixo peso (<2,5kg) e os potenciais fatores associados

| id | age | mwt | race | smoke | nprem | hyper | bwt |
|-----|-----|-----|------|-------|-------|-------|------|
| 1 | 21 | 200 | 2 | 0 | 0 | 0 | 1928 |
| 2 | 16 | 112 | 2 | 0 | 0 | 0 | 3374 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 189 | 26 | 154 | 3 | 0 | 1 | 1 | 2442 |

- Lendo dados no R:

```
> peso=read.table('birthwt.dat',header=TRUE,sep="")
```

- **Dicionário das variáveis:**

age: idade da mãe

mwt: peso da mãe (lbs)

race: raça da mãe (1=Branca, 2=Negra, 3=Outra)

smoke: fumo durante a gravidês (0=Não, 1=Sim)

nprem: Número de partos prematuros

hyper: histórico de hipertensão (0=Não, 1=Sim)

bwt: Peso ao nascer (g)

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 1: Distribuição das mães de recém nascidos segundo raça

| Raça | Frequência absoluta | Frequência relativa |
|--------|---------------------|---------------------|
| Branca | 96 | 0,51 |
| Negra | 26 | 0,14 |
| Outra | 67 | 0,35 |
| Total | 189 | 1 |

Obtendo a distribuição de frequências no R

```
> table(peso$race) #freq absolutas
```

```
> table(peso$race)/length(peso$race) #freq relativas
```

Tabela 2: Distribuição das mães segundo faixa etária

| Idade (anos) | Frequência | | |
|--------------|------------|--------------|---------------|
| | Absoluta | Relativa (%) | Acumulada (%) |
| 10-15 | 6 | 3,17 | 3,17 |
| 15-20 | 63 | 33,33 | 36,50 |
| 20-25 | 66 | 34,92 | 71,42 |
| 25-30 | 34 | 17,99 | 89,41 |
| 30-35 | 17 | 9,00 | 98,41 |
| 35-40 | 2 | 1,06 | 99,47 |
| 40-45 | 1 | 0,53 | 100 |
| Total | 189 | 100 | |

Comandos do R:

```
> h=hist(peso$age,xlab='Idade',ylab='Frequência absoluta',main=")
```

```
> h$counts
```

```
> h$counts/sum(h$counts)*100
```

```
> cumsum(round(h$counts/sum(h$counts)*100,2))
```

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

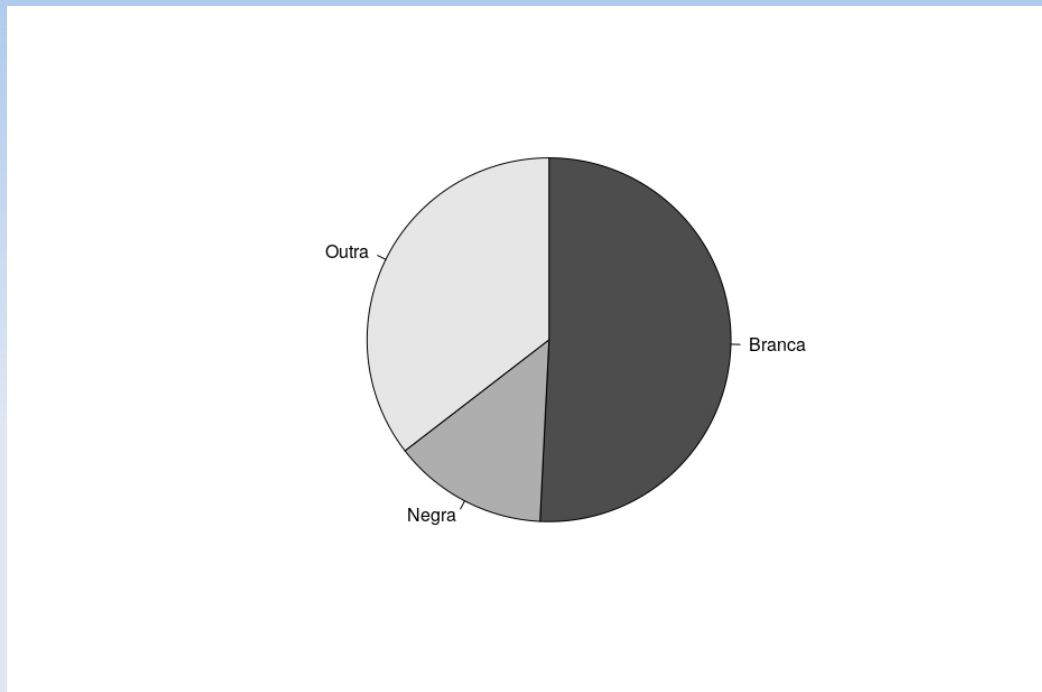
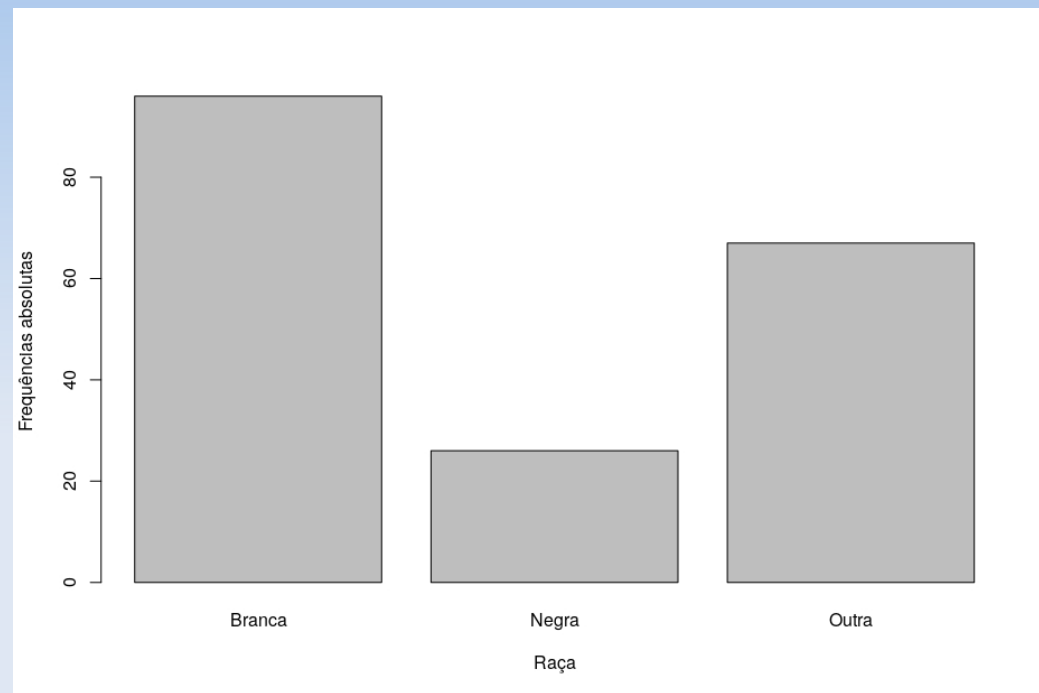
GRÁFICOS

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

Representação gráfica para variáveis categóricas

- Diagrama de barras
- Gráfico de setores

Distribuição das mães de recém nascidos segundo raça



Comandos do R:

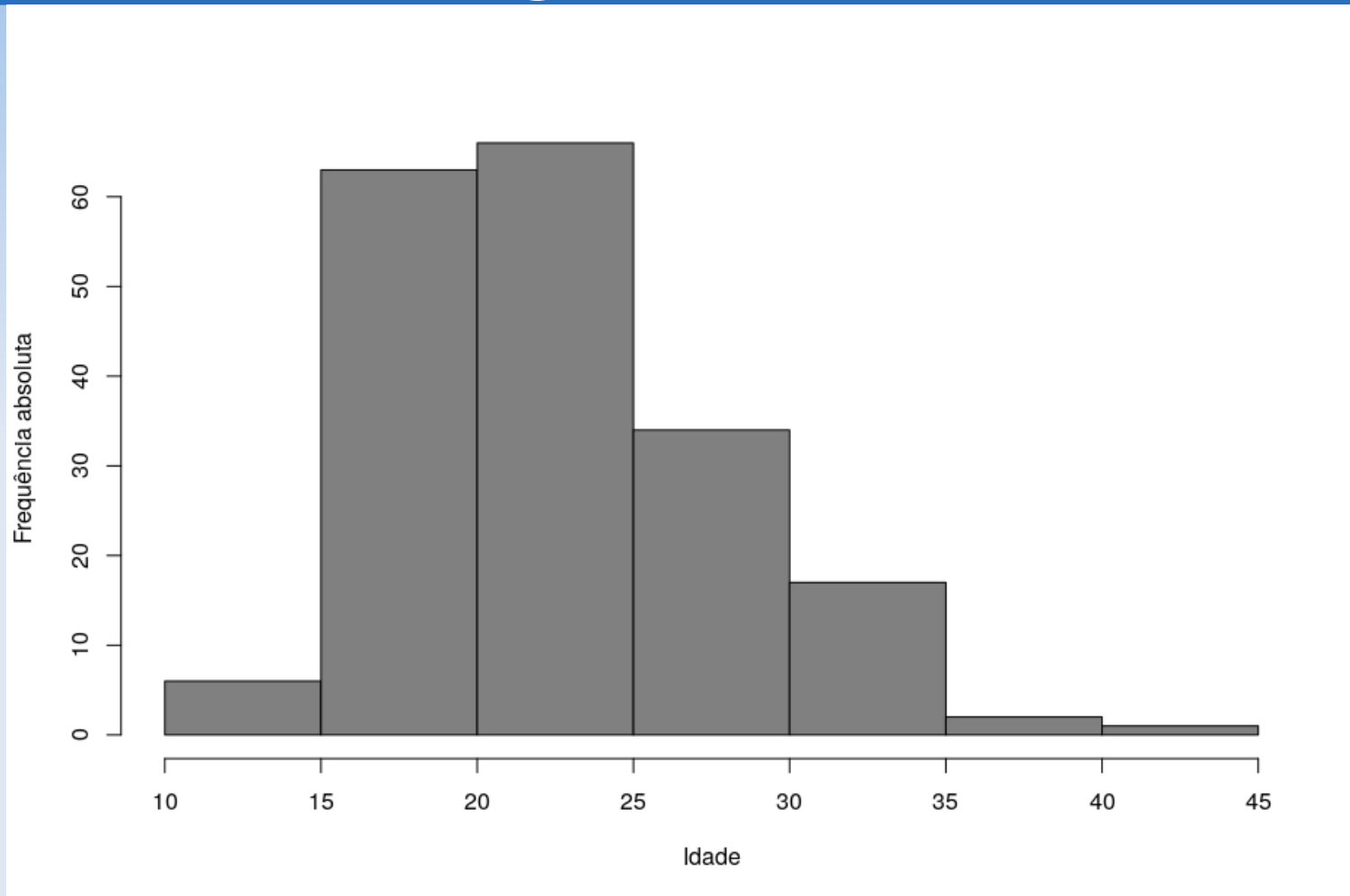
```
> barplot(table(peso$race),names.arg=c("Branca","Negra","Outra"),ylab="Frequências absolutas",xlab="Raça")
```

```
> pie(table(peso$race),labels=c("Branca","Negra","Outra"),clockwise=TRUE,col=grey.colors(3))
```

Representação gráfica de variáveis quantitativas

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

Distribuição de mães de recém nascidos segundo faixa etária



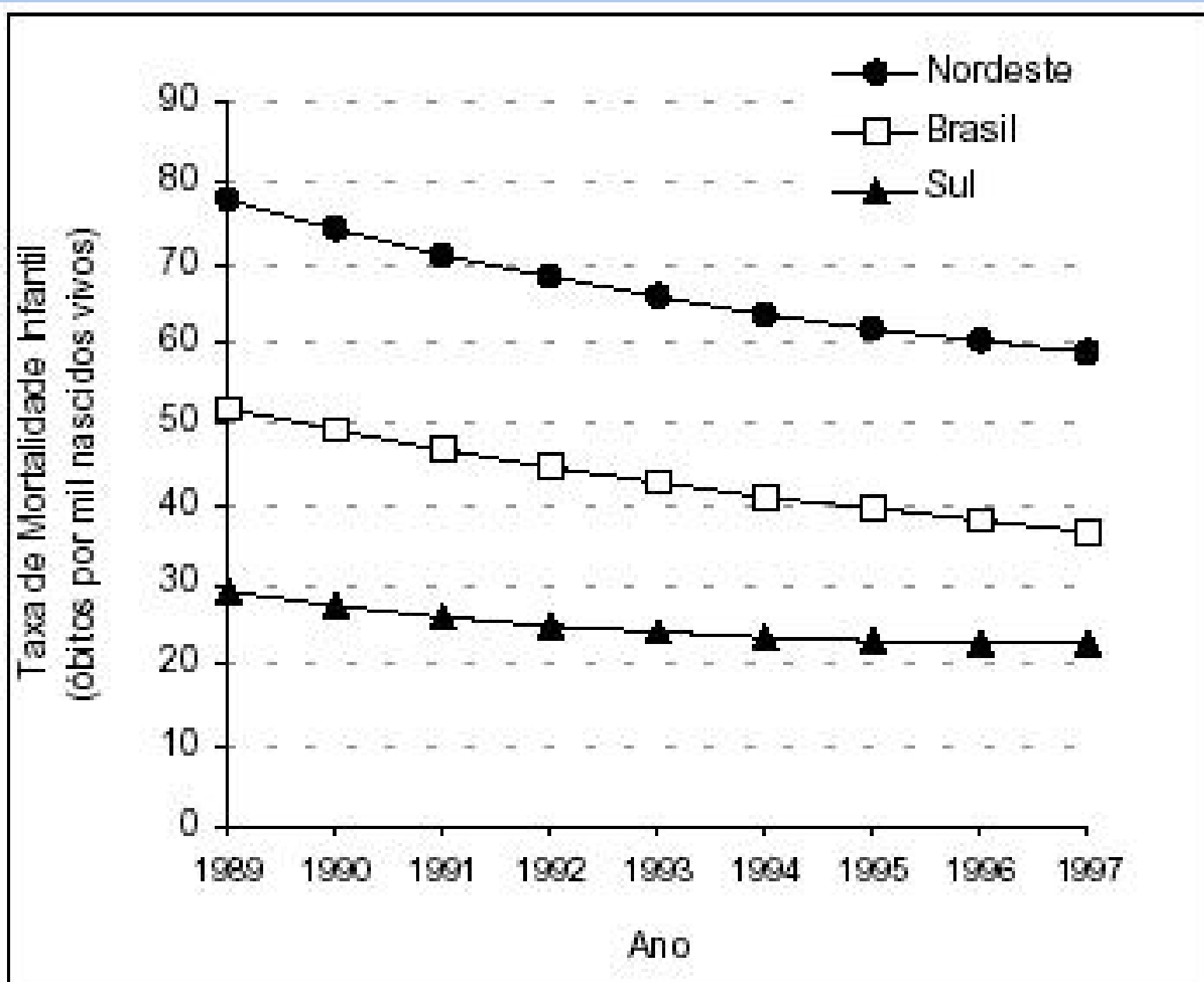
Comando do R:

```
> hist(peso$age,xlab='Idade',ylab='Frequência absoluta',main='',col=grey(.5))
```


Representação gráfica de dados temporais

- Dados coletados ao longo do tempo são comuns em pesquisas médicas
- **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: Recém nascidos

- 39,2% das mães fumaram durante a gravidez
- 6,3% eram hipertensas

Comandos do R:

```
> round(table(peso$smoke)/length(peso$smoke)*100,1)
```

```
> round(table(peso$hyper)/length(peso$hyper)*100,1)
```

Dados Quantitativos

- Para resumir numericamente dados quantitativos escolhemos medidas de:

- **Localção (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 localção e dispersão:

| Localção | 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

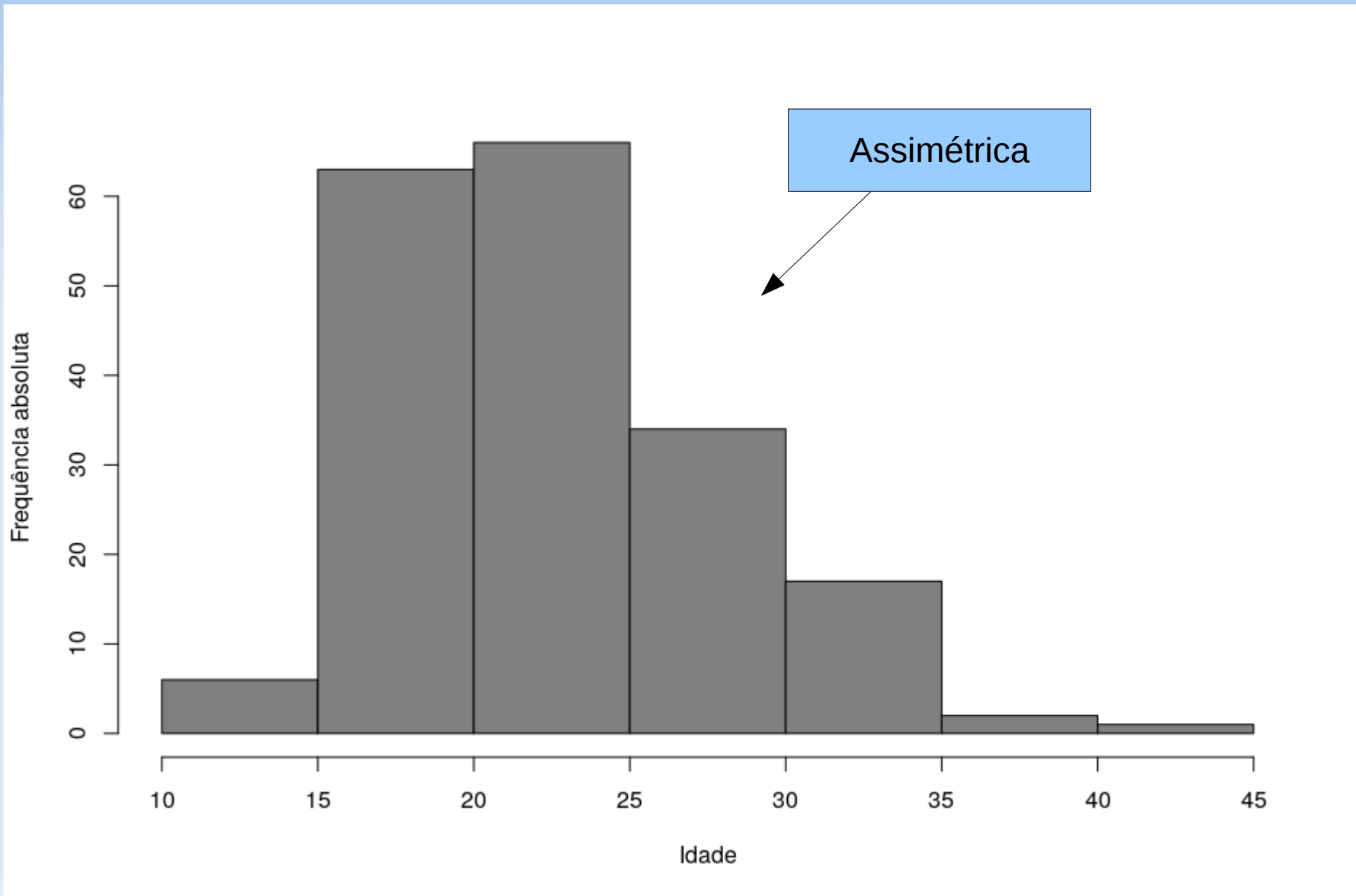
Distribuição de mães de recém nascidos segundo faixa etária

| Idade (anos) | Frequência | | |
|--------------|------------|--------------|---------------|
| | Absoluta | Relativa (%) | Acumulada (%) |
| 10-15 | 6 | 3,17 | 3,17 |
| 15-20 | 63 | 33,33 | 36,50 |
| 20-25 | 66 | 34,92 | 71,42 |
| 25-30 | 34 | 17,99 | 89,41 |
| 30-35 | 17 | 9,00 | 98,41 |
| 35-40 | 2 | 1,06 | 99,47 |
| 40-45 | 1 | 0,53 | 100 |
| Total | 189 | 100 | |

Classe modal



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, $\frac{1}{2}$ dados tem valores maiores do que a mediana, $\frac{1}{2}$ dados tem valores menores do que a mediana.
- **Quartis inferior e superior (Q1 e Q3):** valores baixo dos quais caem $\frac{1}{4}$ e $\frac{3}{4}$ dos dados.
- **5 números sumários (MQMQM):** Min, Q1, Mediana, Q3, Max
- **Amplitude Inter-Quartis:** $AIQ=Q3-Q1$

Usando o R

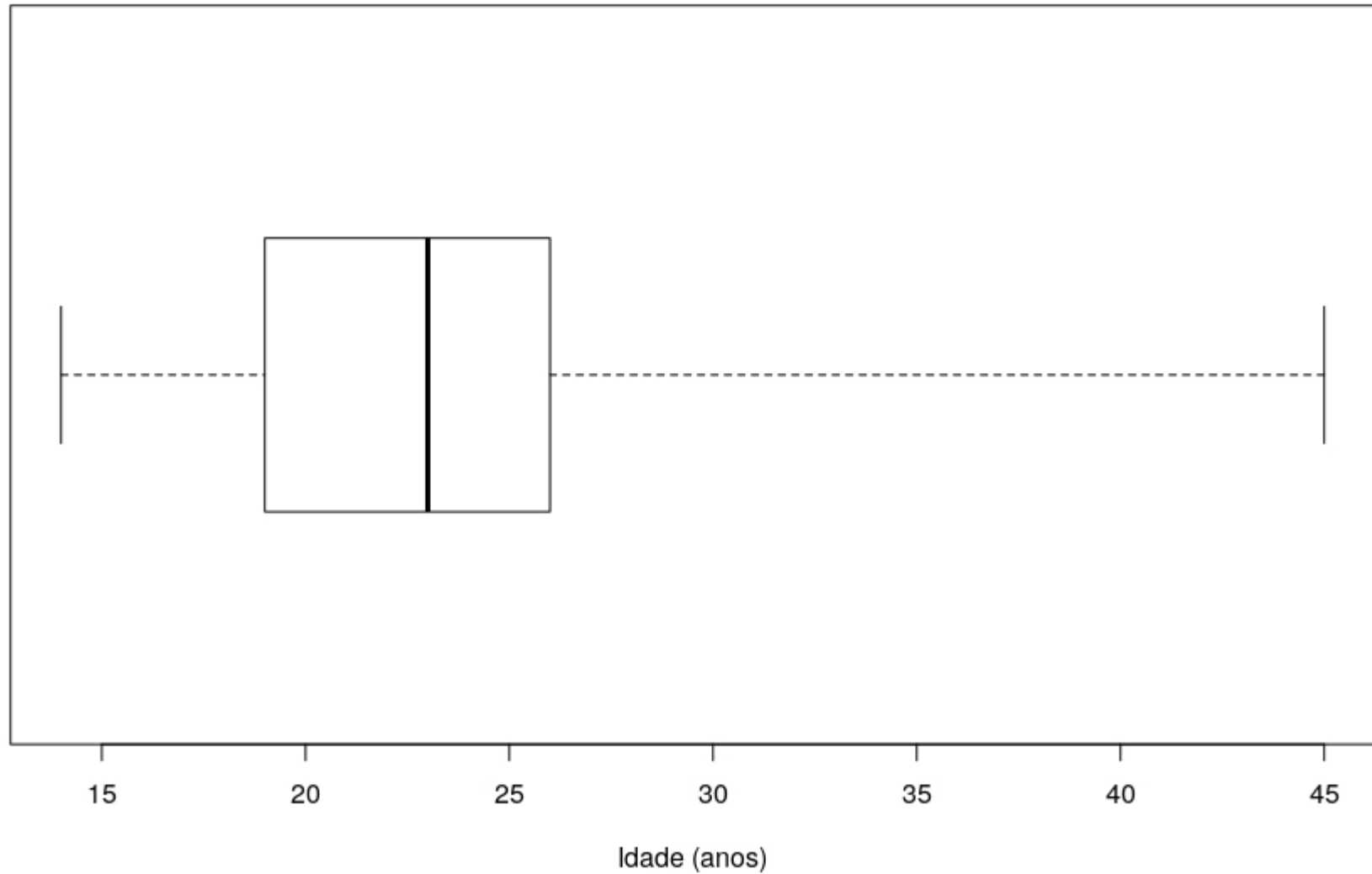
```
> summary(peso$age)
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|-------|---------|--------|-------|---------|-------|
| 14.00 | 19.00 | 23.00 | 23.24 | 26.00 | 45.00 |

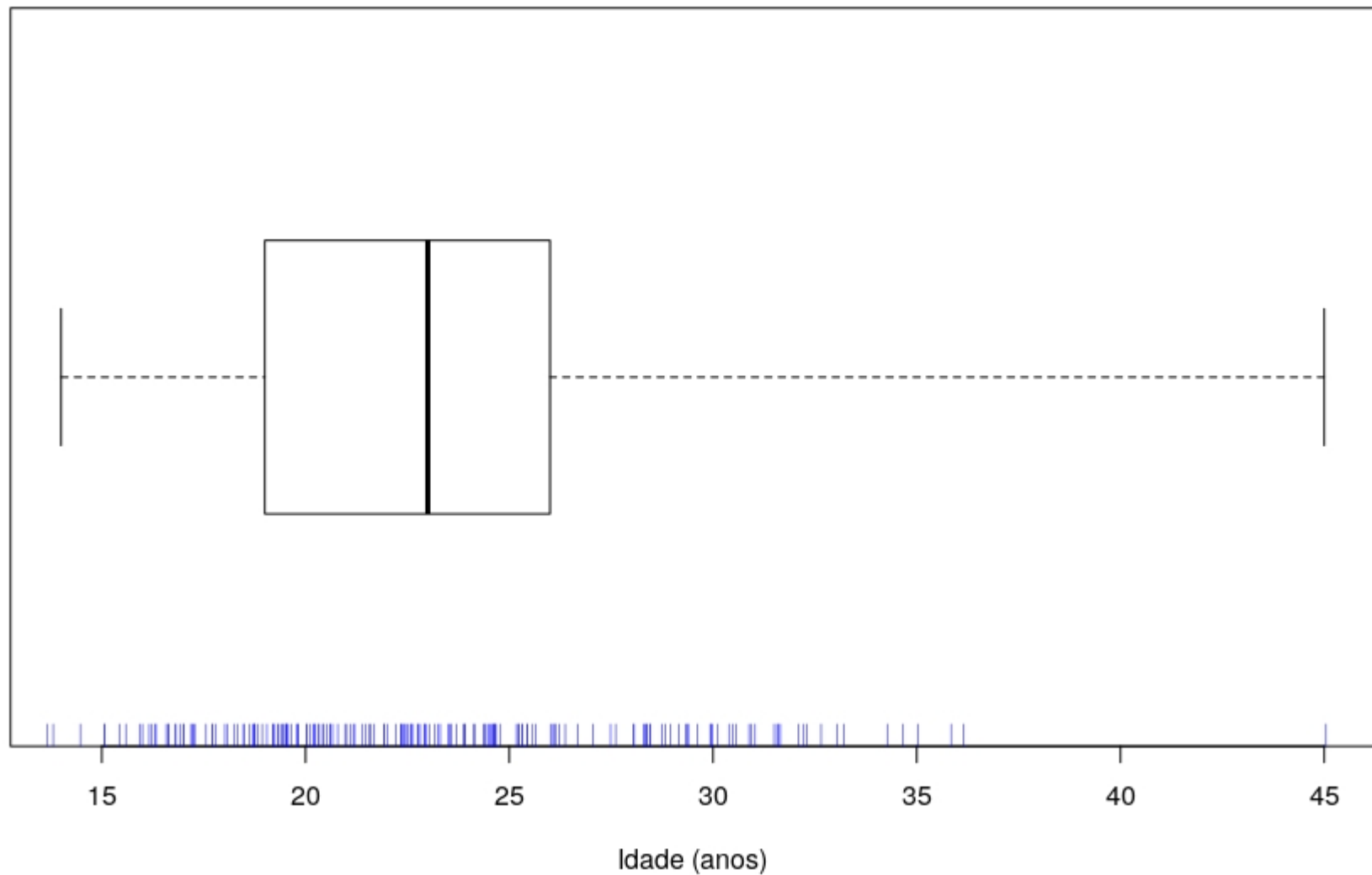
```
> boxplot(peso$age, range=0,xlab='Idade  
(anos)',horizontal=TRUE)
```

```
> rug(jitter(peso$age,amount=0.5), col='blue')
```

Boxplot das idades



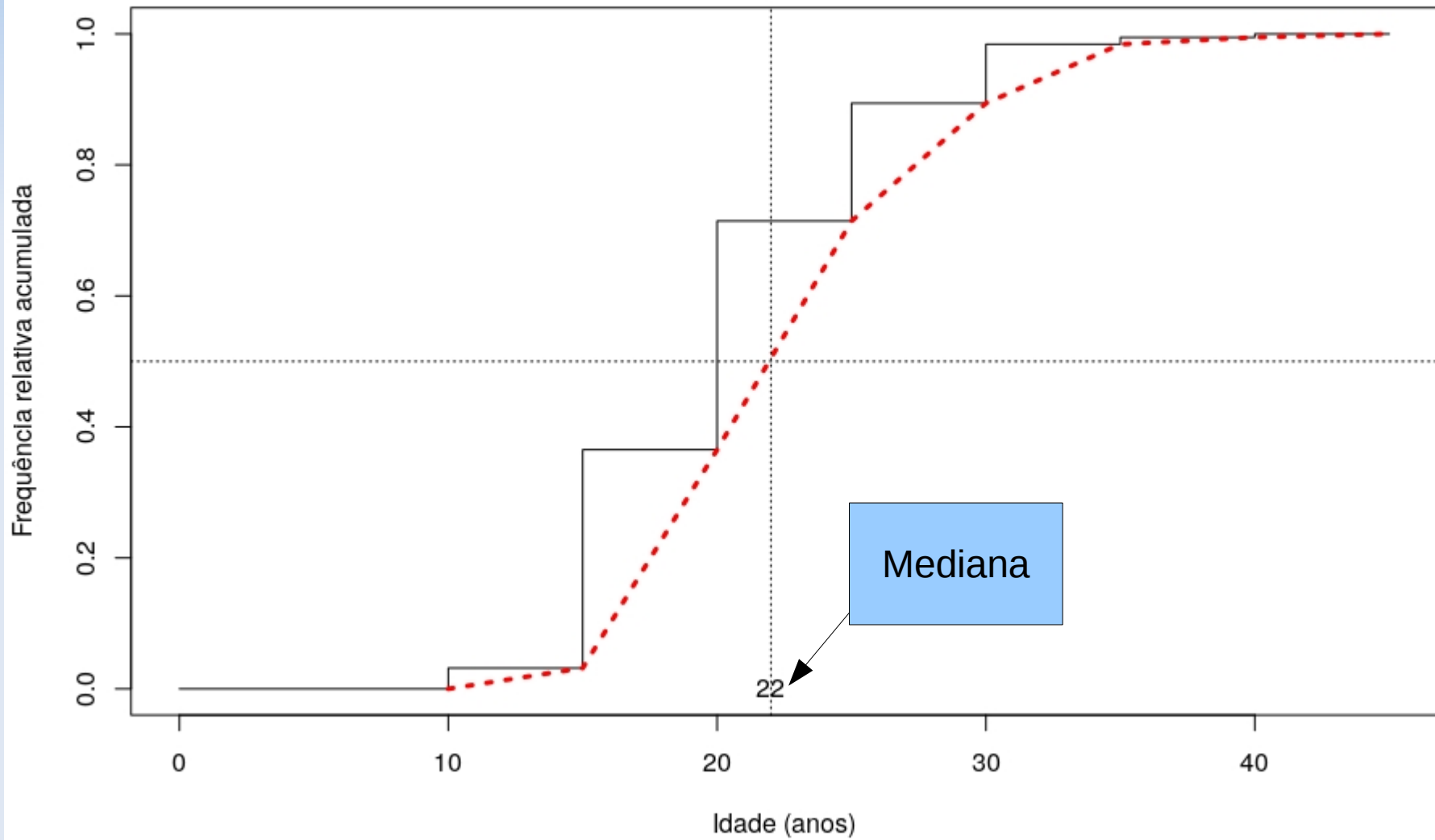
Boxplot das idades



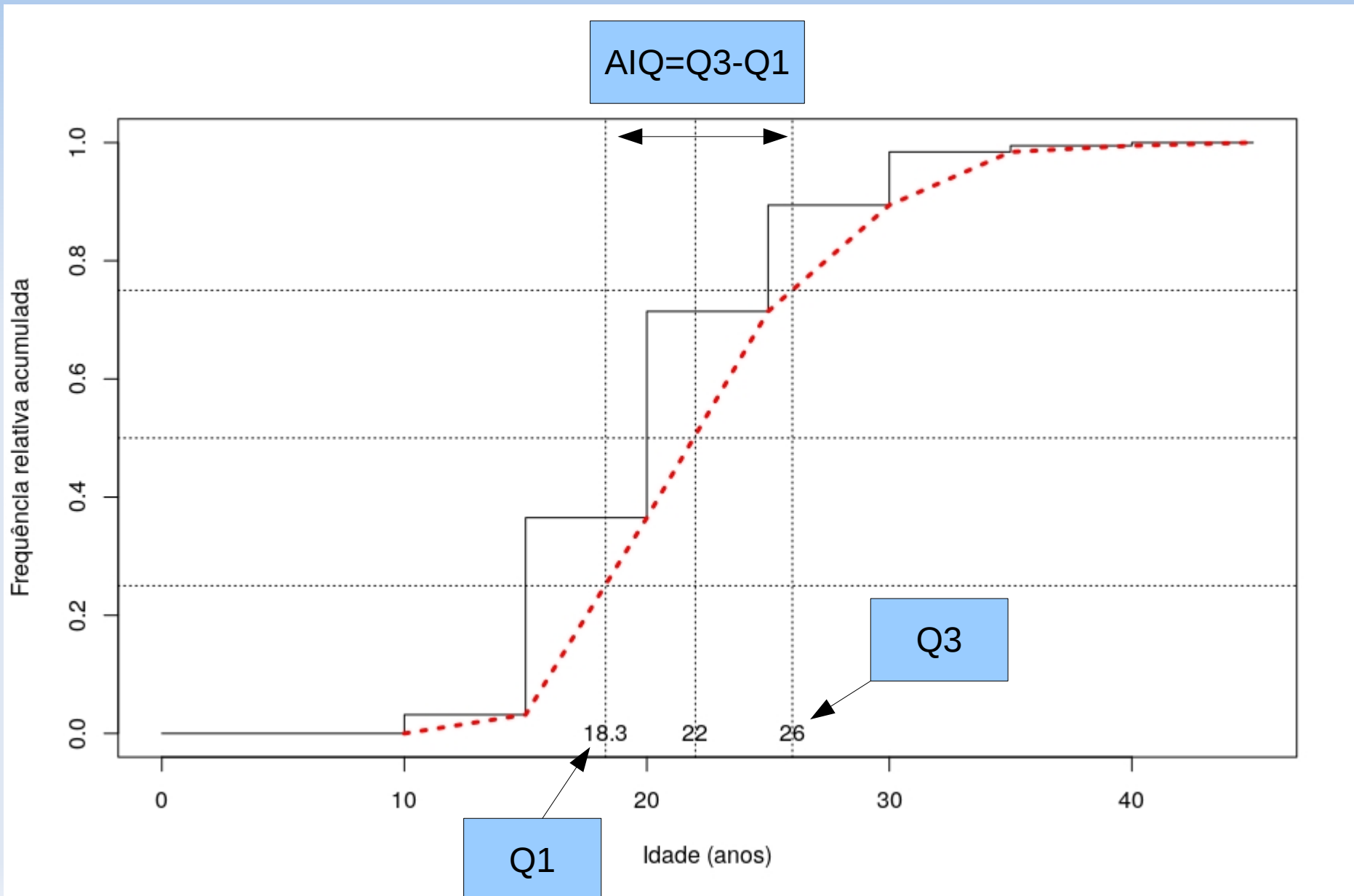
Ogiva

- 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

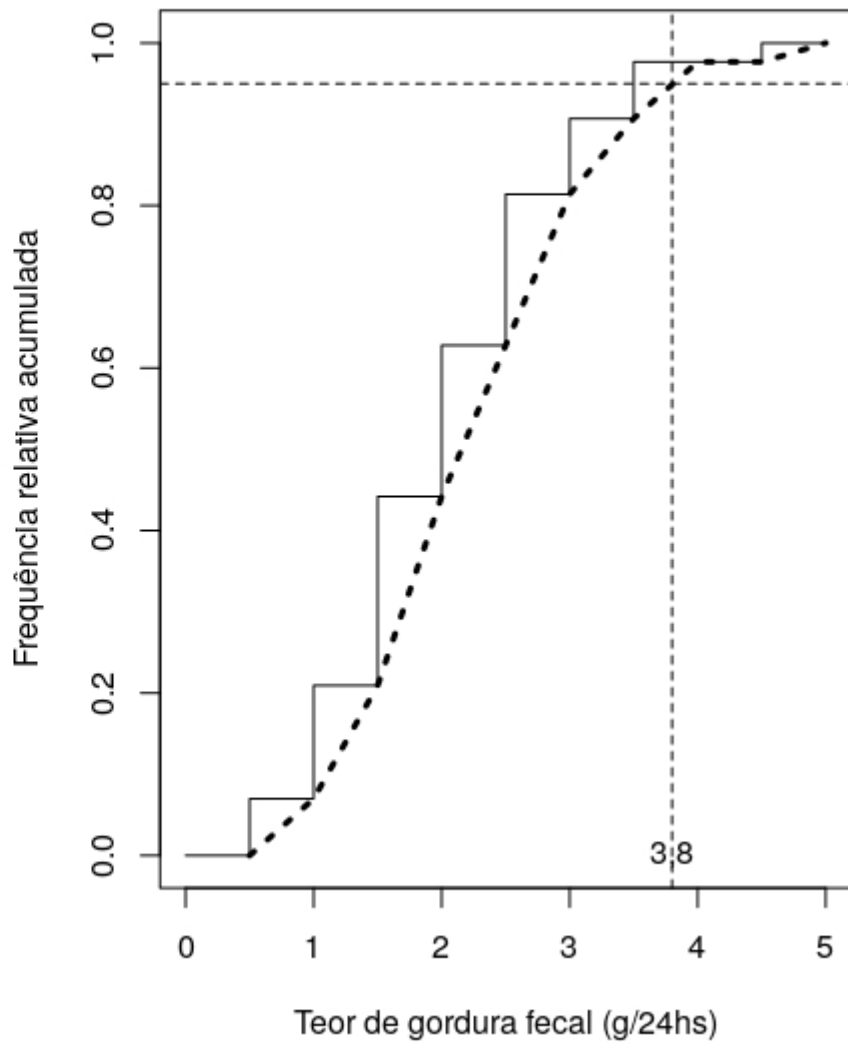
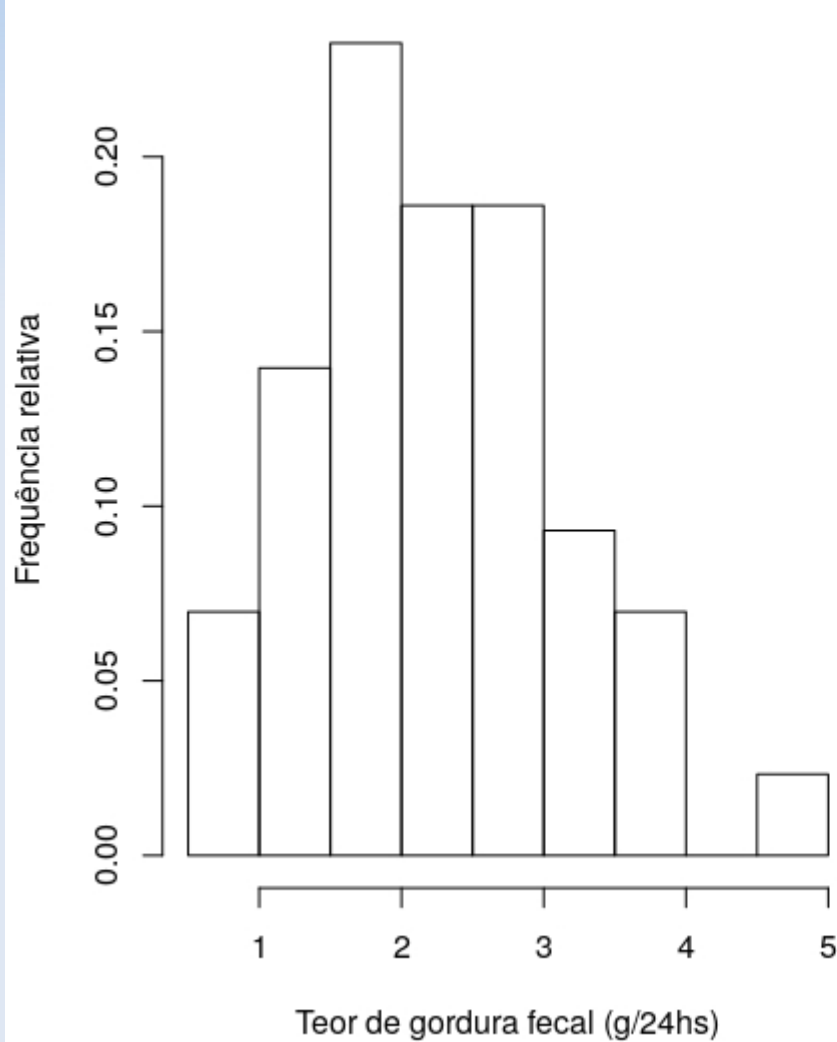
- **Dosagem de gordura:** útil no diagnóstico e acompanhamento da síndrome da má absorção - quando se tem a síndrome tem-se um aumento do teor de gordura fecal.
- Até 1984 não existia um padrão de referência para crianças brasileiras.
- Prof. Francisco Penna (titular de pediatria da UFMG) examinou 43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

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

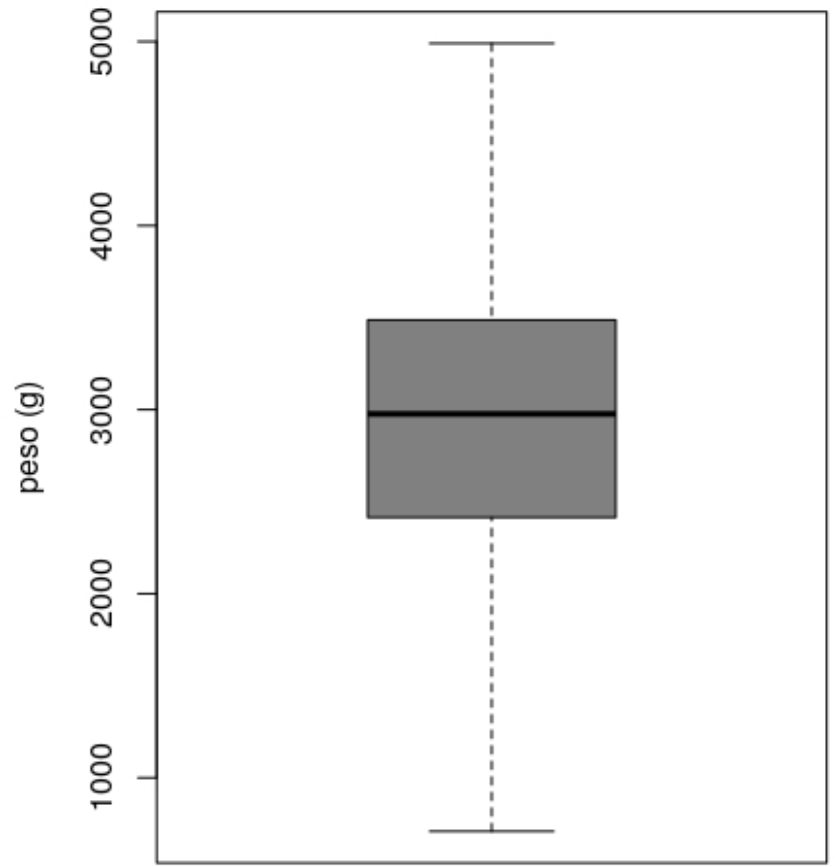
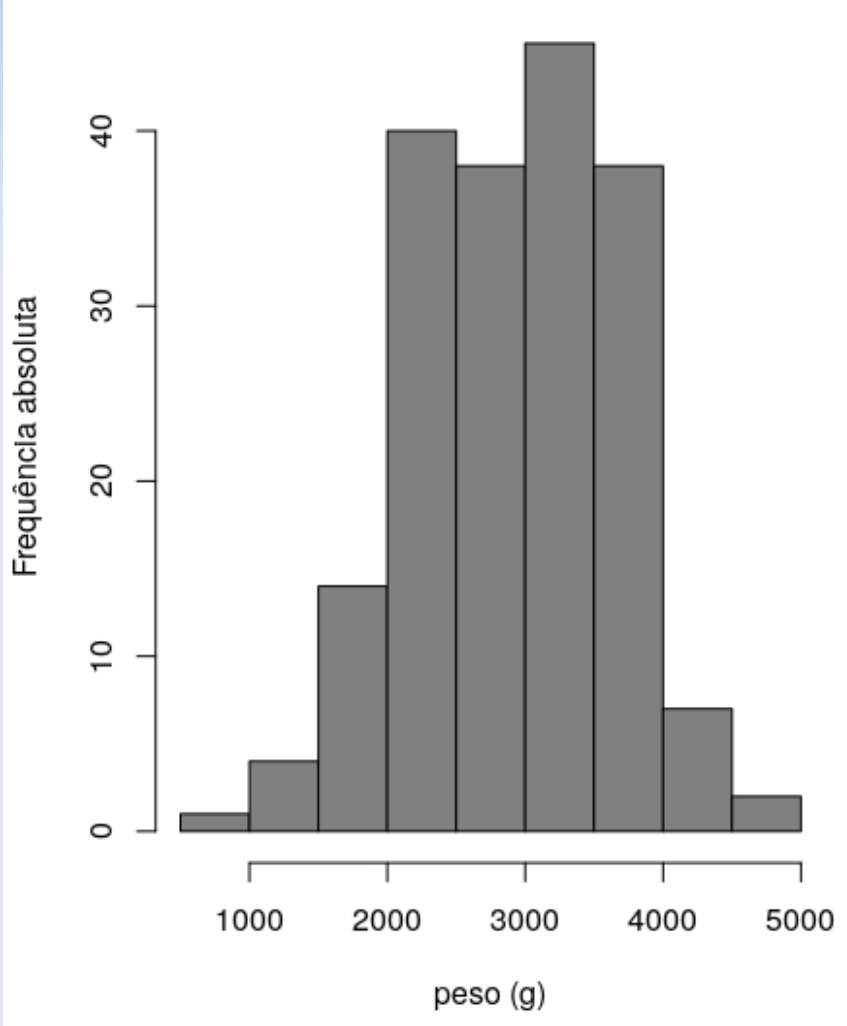
Exemplo: Tentativas de suicídio (cont.)

```
> summary(peso$bwt)
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|------|---------|------|
| 709 | 2414 | 2977 | 2945 | 3487 | 4990 |



Assimetria?



Medidas de variabilidade

- Amplitude total

$$A = \text{Máx} - \text{Min}$$

- Exemplo: Amplitude das idades = $45 - 14 = 31$

É uma boa medida de variabilidade?

Medidas de variabilidade

- **Amplitude total**

$$A = \text{Máx} - \text{Min}$$

- Exemplo: Amplitude das idades = $45 - 14 = 31$

É uma boa medida de variabilidade?

Não utiliza todas as observações.

Medidas de variabilidade

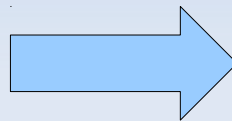
- Considere os conjuntos:

- $A = \{3, 4, 5, 6, 7\}$

- $B = \{1, 3, 5, 7, 9\}$

- $C = \{5, 5, 5, 5, 5\}$

- $D = \{3, 5, 5, 7\}$



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 | 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 |

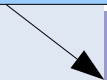
Desvio quadrático médio

| 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

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

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

- n: total amostral

Variância amostral

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 unidade quadrática.
- Para que a dispersão tenha a mesma unidade de medida dos dados originais calculamos a raiz quadrada da variância.

$$\text{Desvio-padrão} = \sqrt{\text{variância}}$$

Notação:

- 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: pesos de recém nascidos

- Pesos de recém nascidos vs tabagismo durante a gravidez

```
> by(peso$bwt,peso$smoke,mean)
```

```
> by(peso$bwt,peso$smoke,sd)
```

| | Tabagismo | |
|---------------|-----------|--------|
| Peso | Sim | Não |
| Média | 2771,9 | 3055,7 |
| Desvio-padrão | 659,6 | 752,7 |

Exemplo: pesos de recém nascidos

- Pesos de recém nascidos vs tabagismo durante a gravidez

```
> by(peso$bwt,peso$smoke,mean)
```

```
> by(peso$bwt,peso$smoke,sd)
```

| | Tabagismo | |
|---------------|-----------|--------|
| Peso | Sim | Não |
| Média | 2771,9 | 3055,7 |
| Desvio-padrão | 659,6 | 752,7 |

- As mães fumantes tem bebês com pesos mais homogêneos do que os bebês de mães não fumantes?

Exemplo: pesos de recém nascidos

- Pesos de recém nascidos vs tabagismo durante a gravidez

```
> by(peso$bwt,peso$smoke,mean)
```

```
> by(peso$bwt,peso$smoke,sd)
```

- Coeficiente de variação

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



Medida de dispersão relativa (pura)

| | Tabagismo | |
|---------------|-----------|--------|
| Peso | Sim | Não |
| Média | 2771,9 | 3055,7 |
| Desvio-padrão | 659,6 | 752,7 |

- As mães fumantes tem bebês com pesos mais homogêneos do que os bebês de mães não fumantes?

Coeficiente de Variação

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



Medida de dispersão relativa (pura)

| | Tabagismo | |
|---------------|-----------|--------|
| Peso | Sim | Não |
| Média | 2771,9 | 3055,7 |
| Desvio-padrão | 659,6 | 752,7 |
| CV (%) | 24 | 25 |

```
> m=by(peso$bwt,peso$smoke,mean)
> s=by(peso$bwt,peso$smoke,sd)
> round(s/m*100)
```

```
peso$smoke: 0
[1] 25
```

```
-----
peso$smoke: 1
[1] 24
```

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 si nas idades ou nos tempos de reação?

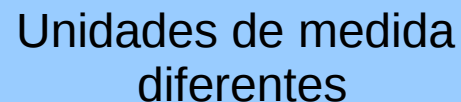
| | Idade | Tempo de reação |
|---------------|-------|-----------------|
| média | 40,23 | 10,22 |
| desvio-padrão | 13,36 | 3,57 |

Exemplo: Teste sorológico

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

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

Unidades de medida diferentes



Coeficiente de Variação

Ex: Os pacientes são mais parecidos entre si 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 |

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



Medida de dispersão relativa (pura)

Escore padronizado

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

| Nota | Média | Desempenho |
|------|-------|------------|
| 7 | 5 | |
| 8 | 6 | |

- 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 | |
| 8 | 6 | 4 | |

Escore padronizado

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

| Nota | Média | Desvio-padrão | Escore Padronizado |
|------|-------|---------------|--------------------|
| 7 | 5 | 2 | 1 |
| 8 | 6 | 4 | 0,5 |



Interpretação?

Usando o R Commander

- Abrir o R e digitar os seguintes comandos:
> `install.packages("Rcmdr",dep=TRUE)`
> `require(Rcmdr)`

Na janela do Rcmdr:

- Selecionar a aba *Dados>Importar arquivos de dados>de arquivo texto, clipboard ou URL...*
- Defina o nome do conjunto de dados: *peso*
- Clique em *OK*
- Na nova janela, selecione o arquivo *birthwt.dat*

Usando o R Commander

- Para visualizar os dados clique na aba: *Ver conjunto de dados*
- Note que todas as colunas foram preenchidas com números inclusive para as variáveis categóricas
- Para dizer ao R quais são as categóricas:
 - Selecione a aba *Dados>Modificação de variáveis no conjunto de dados...>Converter variável numérica para fator...*
 - Selecione todas as variáveis categóricas do banco e na opção *Níveis dos fatores* selecione *Defina nomes dos níveis*

Usando o R Commander

- Para dizer ao R quais são as categóricas:
 - Uma nova janela pergunta: *Sobrescrever variável?*
Selecione a opção: *Sim*
 - Na nova janela digite os nomes dos níveis para a variável
- Fazendo resumos numéricos dos dados
 - Selecione aba: *Estatísticas>Resumos>Conjunto de dados ativo*
- Fazendo gráficos
 - Selecione *Gráficos>Gráfico de Barras*
 - Escolha a variável e clique em *OK*
 - Selecione *Gráficos>Histograma*
 - Escolha a variável e clique em *OK*