

5. Medidas de dispersão

2012



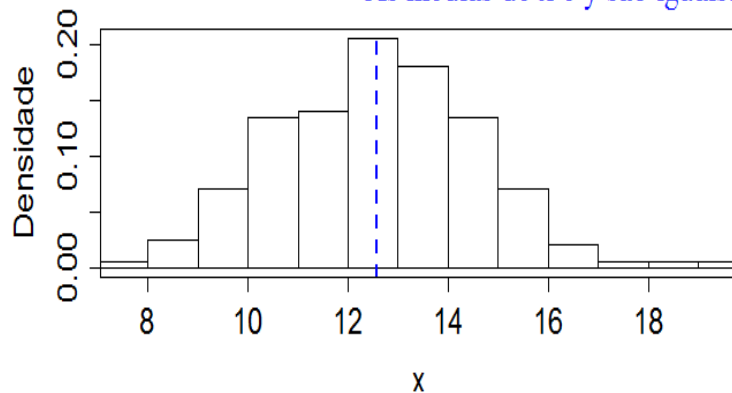
Chamadas de medidas de **variabilidade** (*dispersion* ou *variability*).

Quantificação das **diferenças** entre os valores x_1, x_2, \dots, x_n .

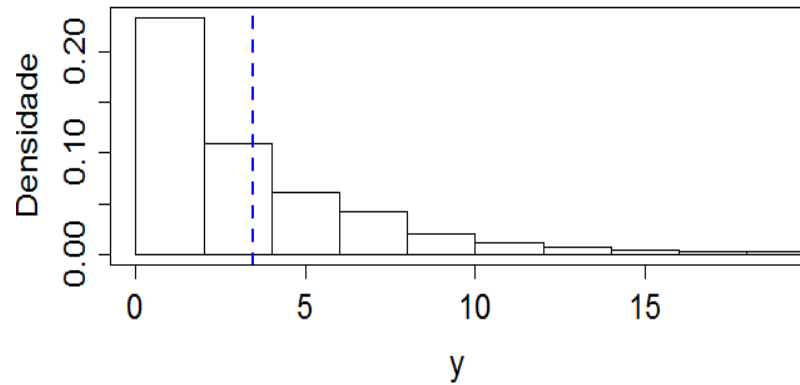
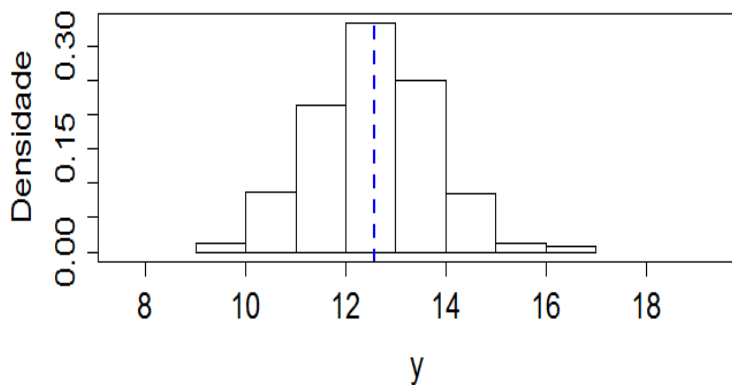
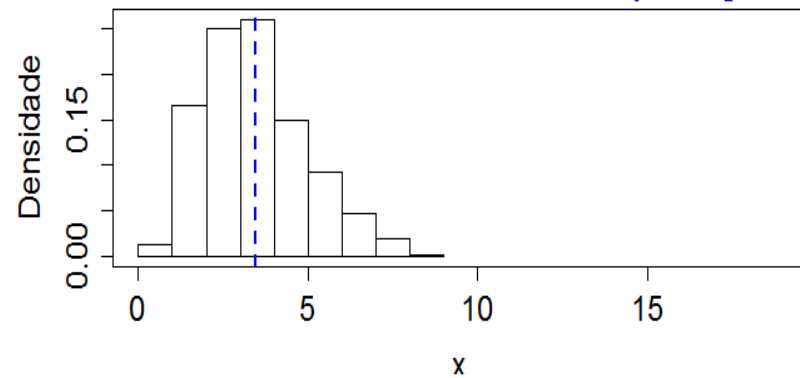
Dispersão e **concentração** (ou **precisão**) são conceitos **opostos**.

Redução (**drástica**) de **n** observações a **um só** valor.

As médias de x e y são iguais.



As médias de x e y são iguais.



5.1. Amplitude (*range*)

Medida de variabilidade entre os **extremos**.

Dados ordenados: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

$$A = x_{(n)} - x_{(1)} = \text{MAX} - \text{min.}$$

Propriedades. (1) $A \geq 0$. (2) $A = 0$ se, e somente se, $x_1 = x_2 = \dots = x_n$.

5.2. Amplitude interquartil (*interquartile range*)

$$d_q = Q_3 - Q_1.$$

d_q é **mais resistente** do que A .

Exercícios. (1) Apresente a curva de sensibilidade (CS) de A .

(2) Qual a forma da CS de d_q ?



Valor atípico (*outlier*)

Valor extremo, espúrio, aberrante, estranho, discrepante,...

Observação **afastada** do **restante** dos dados.

Critérios:

$x_i < Q_1 - 3d_q$ ou $x_i > Q_3 + 3d_q$: valor atípico **severo**.

$Q_1 - 3d_q < x_i < Q_1 - 1,5d_q$ ou

$Q_3 + 1,5d_q < x_i < Q_3 + 3d_q$: valor atípico **moderado**.

$Q_1 - 3d_q$: **barreira externa inferior**.

$Q_1 + 3d_q$: **barreira externa superior**.

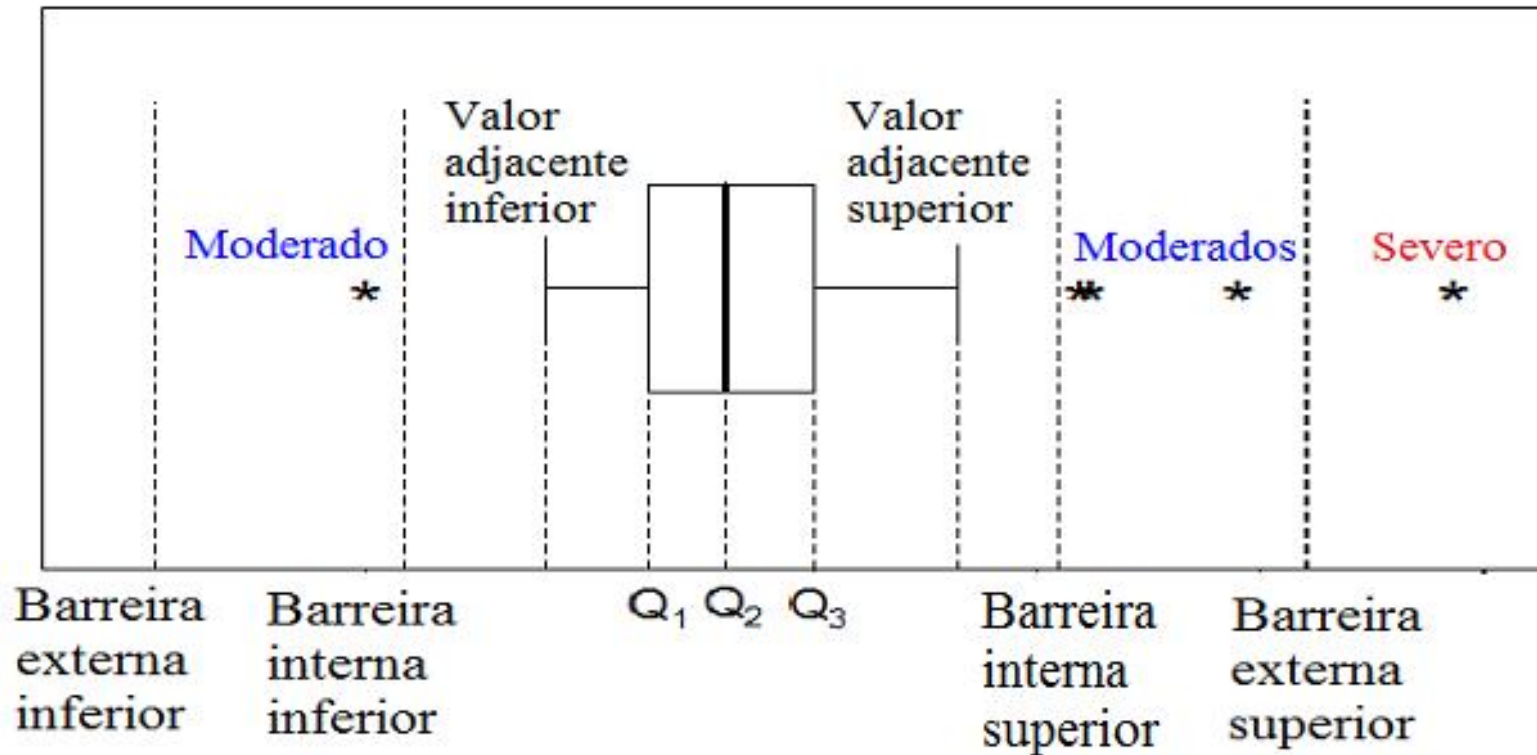
$Q_1 - 1,5d_q$: **barreira interna inferior**.

$Q_1 + 1,5d_q$: **barreira interna superior**.



Gráfico de caixa (*box plot*)

Gráfico caixa-de-bigodes (*box-and-whisker plot*)



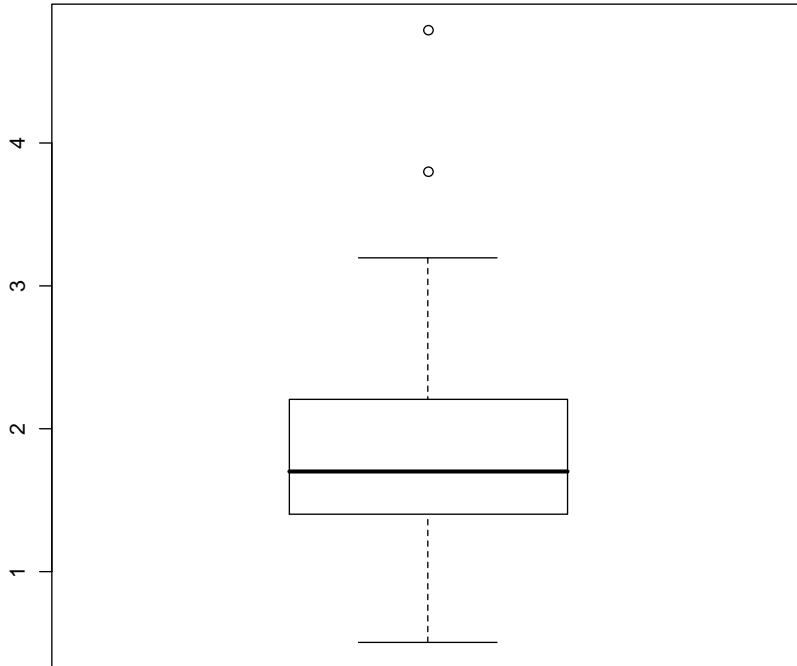
Valor adjacente inferior: menor valor no conjunto de dados que não é extremo (pode ser igual a $x_{(1)}$).

Valor adjacente superior: maior valor no conjunto de dados que não é extremo (pode ser igual a $x_{(n)}$).

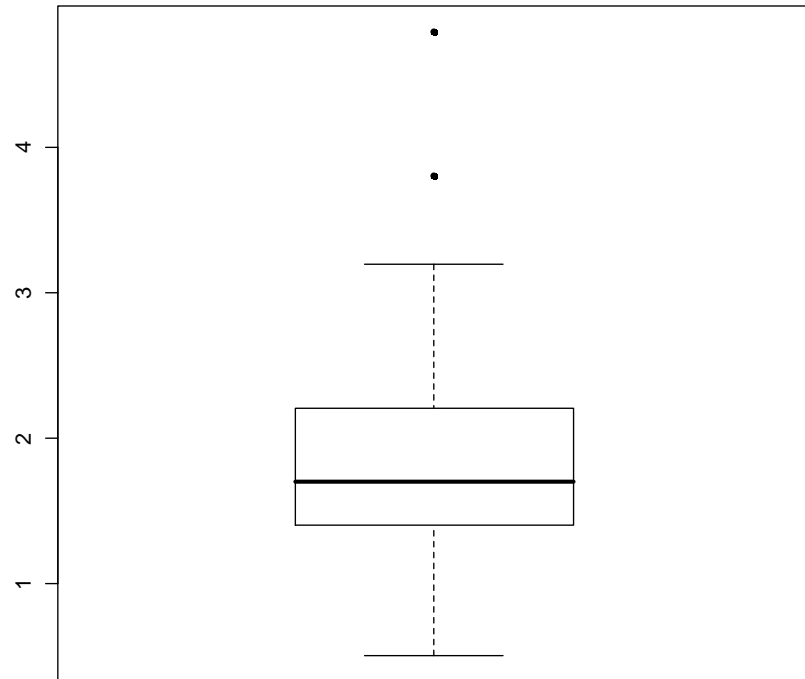
Gráfico de caixa

```
x = c(1.5, 1.9, 1.7, 1.6, 3.8, 1.3, 2.2, 1.8, 1.3, 0.5, 1.6, 1.4, 1.7, 1.7,  
1.9, 0.7, 2.2, 2.3, 2.4, 2.3, 1.8, 2.7, 1.3, 1.7, 2.0, 1.1, 2.1, 1.6, 1.3,  
2.2, 1.5, 2.3, 1.1, 1.8, 1.2, 2.0, 1.5, 1.5, 2.6, 1.6, 1.4, 2.2, 1.5, 1.2,  
2.0, 1.3, 2.6, 1.9, 1.3, 2.4, 3.2, 1.9, 4.8)
```

```
> boxplot(x)
```



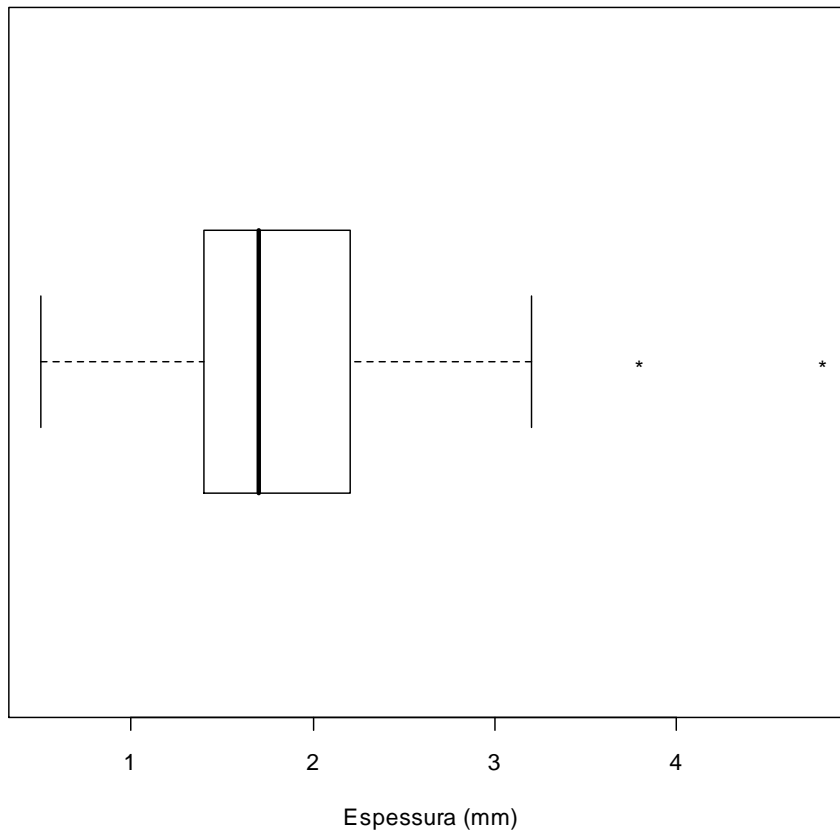
```
> boxplot(x, pch = 20)
```



Obs. Na construção do gráfico de caixas, quanto maior for n, melhor.

Gráfico de caixa

```
> boxplot(x, pch = "*",  
horizontal = TRUE, xlab =  
"Espessura (mm)")
```



```
> bx = boxplot(x, plot = FALSE)
```

```
> names(bx)
```

```
[1] "stats" "n" "conf"  
"out" "group" "names"
```

`bx$stats`: valor adjacente inferior, Q_1 , Q_2 , Q_3 e valor adjacente superior.

`bx$n`: número de observações.

`bx$out`: observações extremas.

```
      [,1]  
> bx$stats  
[1,] 0.5  
[2,] 1.4  
[3,] 1.7  
[4,] 2.2  
[5,] 3.2
```

```
> class(bx$stats)
```

```
[1] "matrix"
```



Gráfico de caixa

```
> boxplot(x, pch = "*", horizontal = TRUE, xlab = "Espessura (mm)")  
> identify(box$out, rep(1, length(box$out)), match(box$out, x))
```

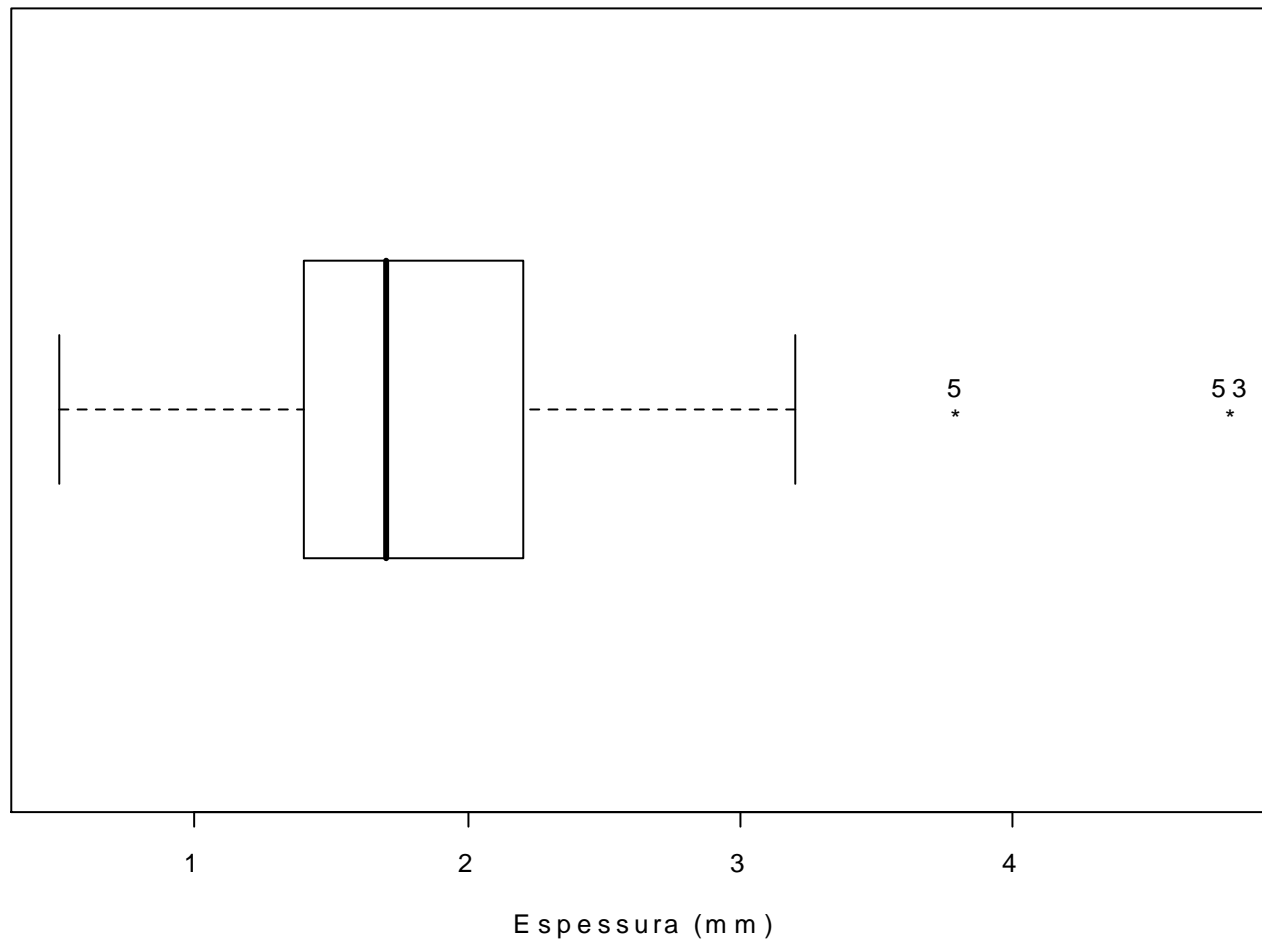
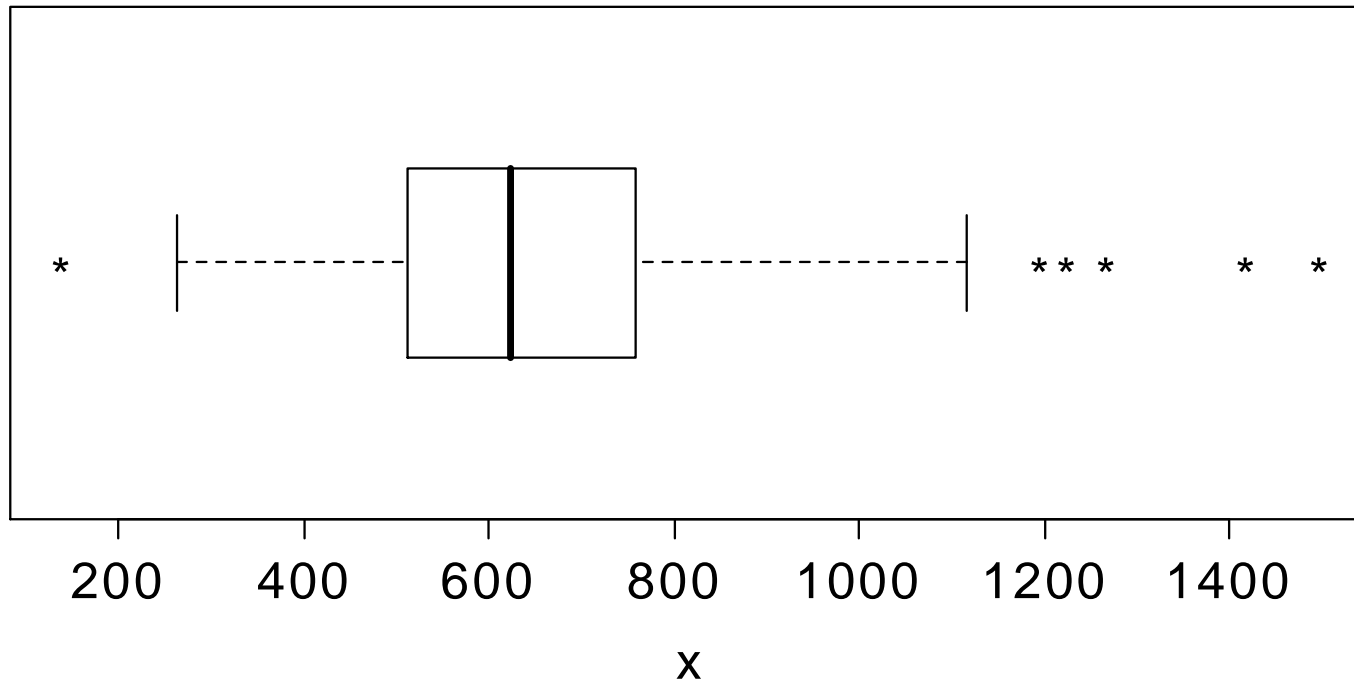


Gráfico de caixa

O que é possível observar em um gráfico de caixa?

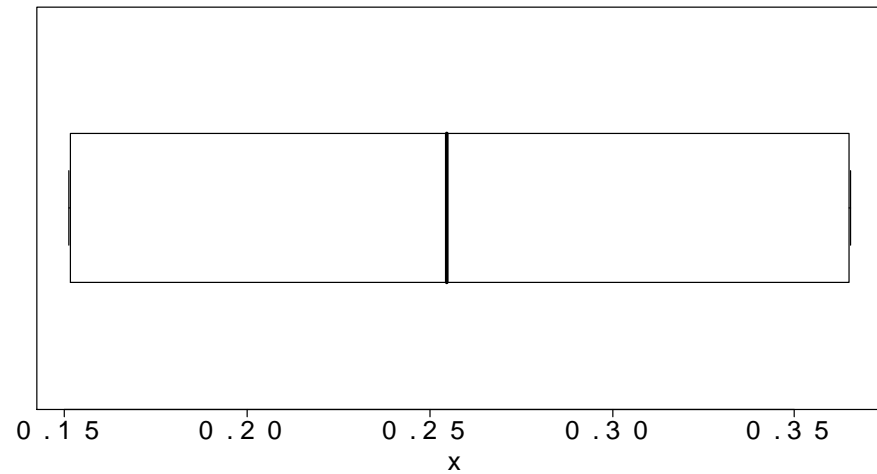
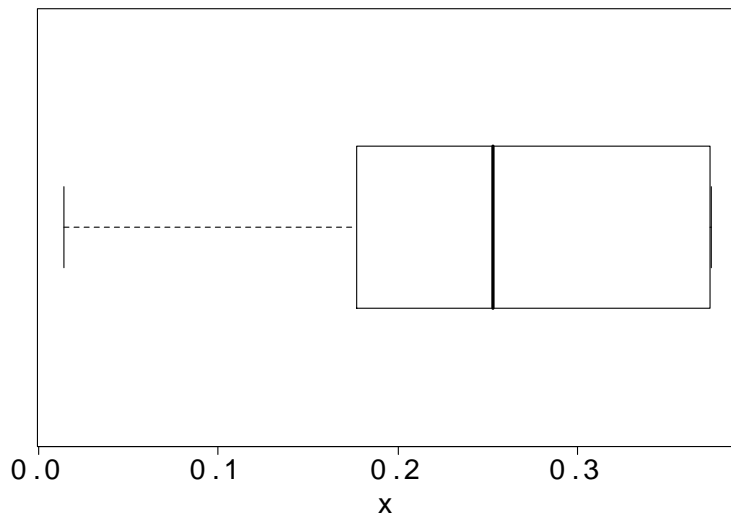
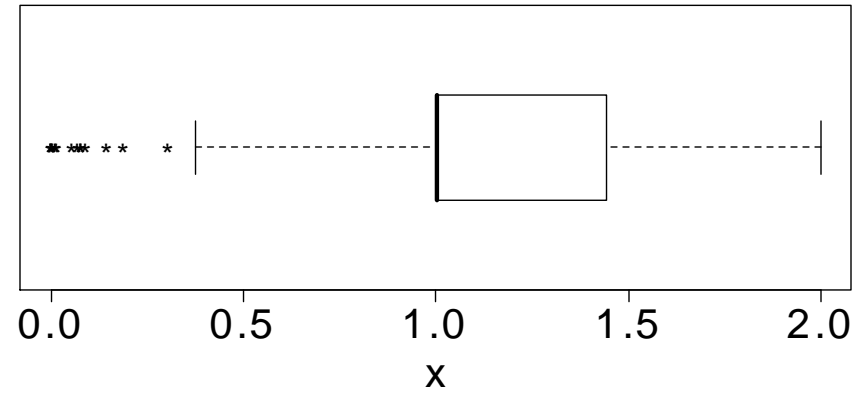
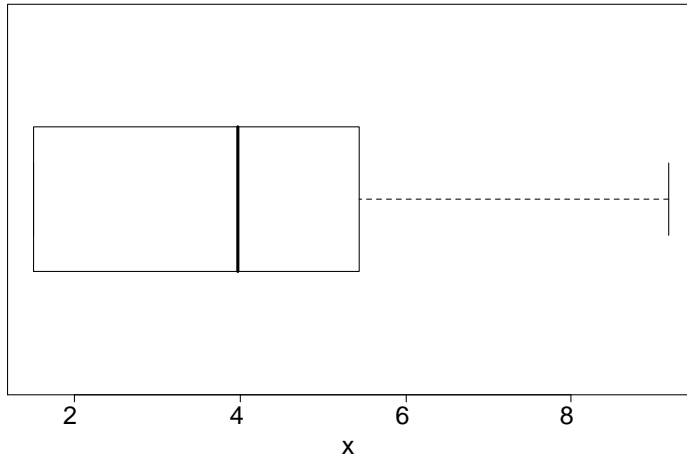


Medida de **posição** ($M = Q_2$). Medida de **dispersão** ($d_q = Q_3 - Q_1$).

Simetria. Valores **extremos**.



Gráfico de caixa



Exercício. Descreva conjuntos de dados correspondentes a cada um dos gráficos.

5.3. Desvio médio ou desvio absoluto médio (*mean absolute deviation*)

$$dm = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

Obs. (1). A mediana (M) pode ser usada no lugar da média.

(2) Não é uma medida resistente.

5.4. Desvio absoluto mediano (*median absolute deviation*).

$M = \text{mediana}(x_1, x_2, \dots, x_n)$.

$MAD = \text{mediana}(|x_1 - M|, |x_2 - M|, \dots, |x_n - M|)$.

Obs. MAD é uma medida resistente.



5.5. Variância (*variance*)

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

Obs. (1). Unidade de s^2 é a unidade de x^2 .

(2) **Não é** uma medida **resistente**.

(3) **Importante** em Inferência Estatística.

Exercício. Prove que $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.

5.6. Desvio padrão (*standard deviation*)

$$s = \sqrt{s^2} = \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2}.$$

Obs. (1). Unidade de s é a **mesma** unidade de x .

(2) **Não é** uma medida **resistente**.



Propriedades da variância

P1. Se $y_i = a + x_i$, $i = 1, \dots, n$, a um número real, então $s_y^2 = s_x^2$.

P2. Se $y_i = bx_i$, $i = 1, \dots, n$, b um número real, então $s_y^2 = b^2 s_x^2$.

Obs. $s_y = |b| s_x$.

P3. Se $y_i = a + bx_i$, $i = 1, \dots, n$, a e b números reais, então $s_y^2 = b^2 s_x^2$.

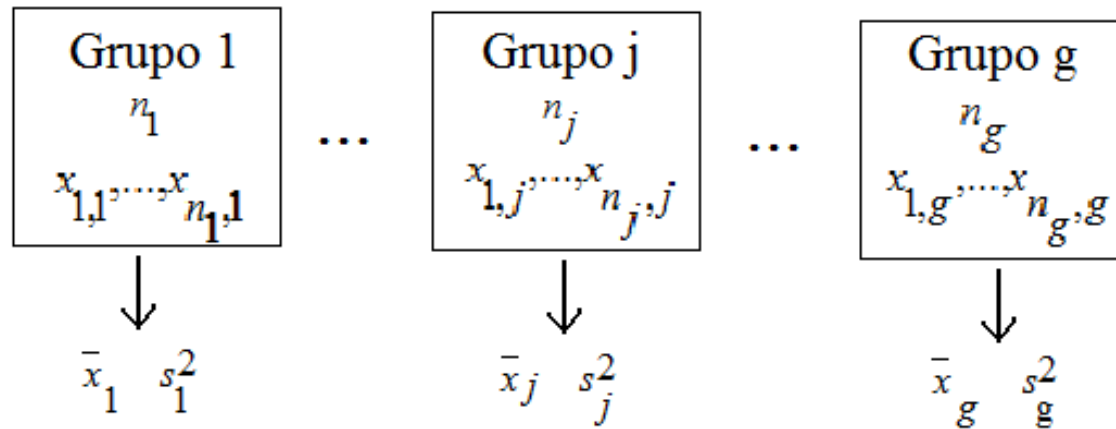
P4. Se as n observações compõem g grupos ($g \geq 2$), cada um com $n_j \geq 2$ observações e $n_1 + n_2 + \dots + n_g = n$, então

$$\begin{aligned}(n-1)s^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{variação total em relação à média}) \\ &= \sum_{j=1}^g (n_j - 1)s_j^2 + \sum_{j=1}^g n_j (\bar{x}_j - \bar{x})^2.\end{aligned}$$

Obs. Variação total = variação intragrupos + variação entre grupos.

Total variation = within groups variation + between groups variation.

Propriedades da variância



$$\bar{x}_j = \frac{1}{n_j} \sum_{m=1}^{n_j} x_{m,j}, \quad j = 1, \dots, g. \quad \bar{x} = \frac{n_1 \bar{x}_1 + \dots + n_g \bar{x}_g}{n_1 + \dots + n_g} = \frac{1}{n} \sum_{j=1}^g n_j \bar{x}_j.$$

$$s_j^2 = \frac{1}{n_j - 1} \sum_{m=1}^{n_j} (x_{j,m} - \bar{x}_j)^2, \quad j = 1, \dots, g.$$



Exemplo – dados na lâmina 6

```
x = c(1.5, 1.9, 1.7, 1.6, 3.8, 1.3, 2.2, 1.8, 1.3, 0.5, 1.6, 1.4, 1.7, 1.7,  
1.9, 0.7, 2.2, 2.3, 2.4, 2.3, 1.8, 2.7, 1.3, 1.7, 2.0, 1.1, 2.1, 1.6, 1.3,  
2.2, 1.5, 2.3, 1.1, 1.8, 1.2, 2.0, 1.5, 1.5, 2.6, 1.6, 1.4, 2.2, 1.5, 1.2,  
2.0, 1.3, 2.6, 1.9, 1.3, 2.4, 3.2, 1.9, 4.8)
```

```
> var(x)
```

```
[1] 0.5059652
```

```
> sd(x)
```

```
[1] 0.7113123
```

```
> xb = mean(x)
```

```
> (dm = mean(abs(x - xb)))
```

```
[1] 0.498042
```

```
> M = median(x)
```

```
> (MAD = median(abs(x - M)))
```

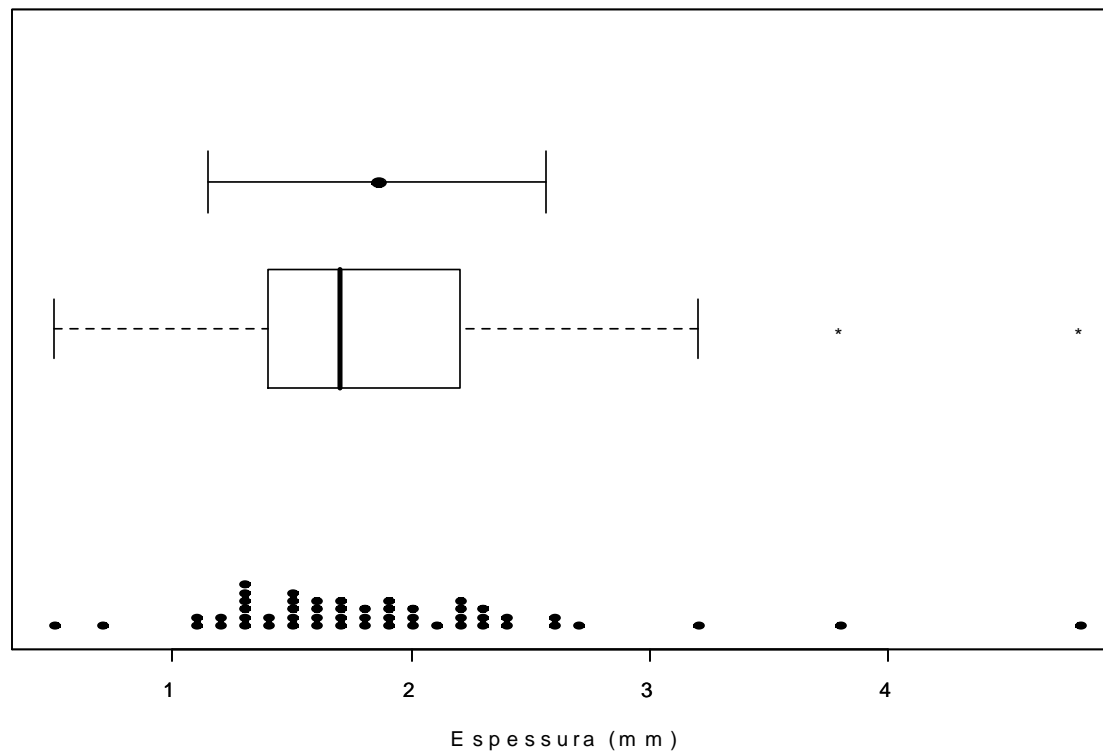
```
[1] 0.4
```

Exercício. Consulte a ajuda da função `mad` (`? mad`).



Exemplo – dados na lâmina 6

```
> despad = sd(x)
> stripchart(x, method = "stack", pch = 20, xlab = "Espessura (mm)",
at = 0)
> boxplot(x, pch = "*", horizontal = TRUE, at = 1, add = TRUE)
> arrows(xb - despad, 1.5, xb + despad, 1.5, code = 3, angle = 90)
> points(xb, 1.5, pch = 19)
```



5.7. Coeficiente de variação (*coefficient of variation*)

(1) O desvio padrão (s) está vinculado à média.

Dificuldade em comparar desvios padrão se as **médias** são muito diferentes.

(2) A , d_q , dm , MAD , s^2 e s são medidas de dispersão **absolutas**.

Dependem da unidade de medida de x .

Comparações envolvendo duas ou mais **variáveis diferentes** ou medidas em **diferentes escalas** (m e cm, p. ex.) **não são possíveis**.

(1) e (2) apontam a conveniência de medidas **relativas**.

$$CV = \frac{s}{|\bar{x}|}, \quad \text{se } |\bar{x}| \neq 0. \quad \text{Pode ser dado em \%}.$$

Propriedades. (1) CV é **adimensional**.

(2) **Não é** uma medida **resistente**.

(3) É **instável** se **média** $\cong 0$.

(4) $0 \leq CV < n^{1/2}$.



5.8. Amplitude studentizada (*Studentized range*)

$$A_s = \frac{A}{s} = \frac{x_{(n)} - x_{(1)}}{s} = \frac{\text{MAX} - \text{min}}{s}. \quad \text{Pode ser dada em \%}.$$

Obs. Dividir pelo desvio padrão significa studentizar (ou padronizar) uma medida.

Propriedades. (1) Não é uma medida resistente.

$$(2) \quad 2\sqrt{\frac{n-1}{n}} \leq A_s \leq \sqrt{2(n-1)}.$$

Obs. Uma medida de dispersão relativa resistente: d_q / M .

Exemplo – dados na lâmina 6

```
> (cv = sd(x) / mean(x))           > (As = (max(x) - min(x)) / sd(x))
[1] 0.3831255                       [1] 6.045165
```

Obs. A função `range` fornece o vetor (min, MAX).

