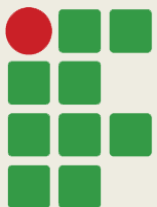
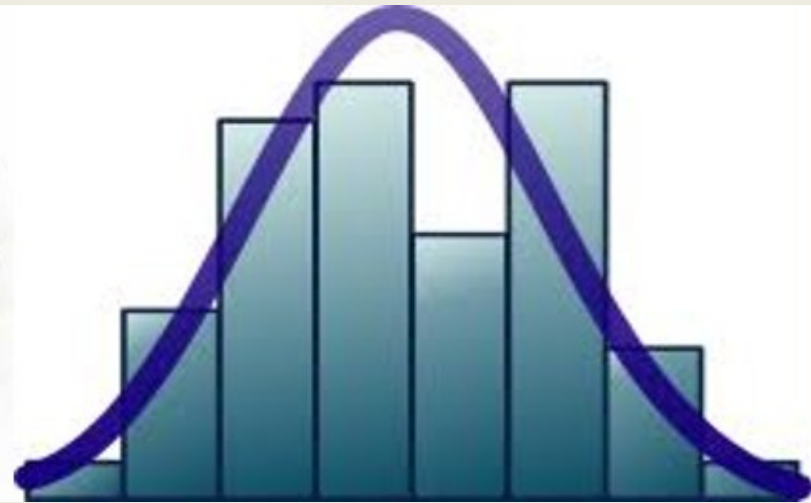


Probabilidade e Estatística



INSTITUTO FEDERAL
Catarinense
Campus Blumenau

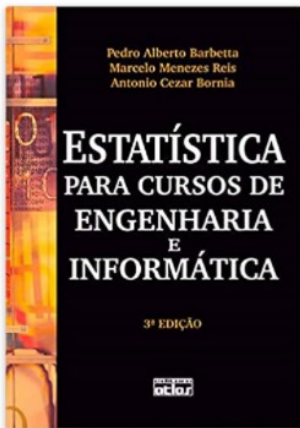
Professor Jeovani Schmitt



Probabilidade e Estatística

Aula 6

Análise Exploratória de Dados (AED)



- ☐ Box Plot e assimetria
- ☐ Exercícios no R
(atividade 3)

Exemplo 1: Notas da P1 de 40 alunos da TURMA A

0	10	10	20	30	40	40	40	40	50
50	50	50	50	60	60	60	60	70	70
70	70	70	70	80	80	80	80	80	80
80	80	90	90	90	90	100	100	100	100

Realize uma análise exploratória das notas.

Aula passada estudamos:

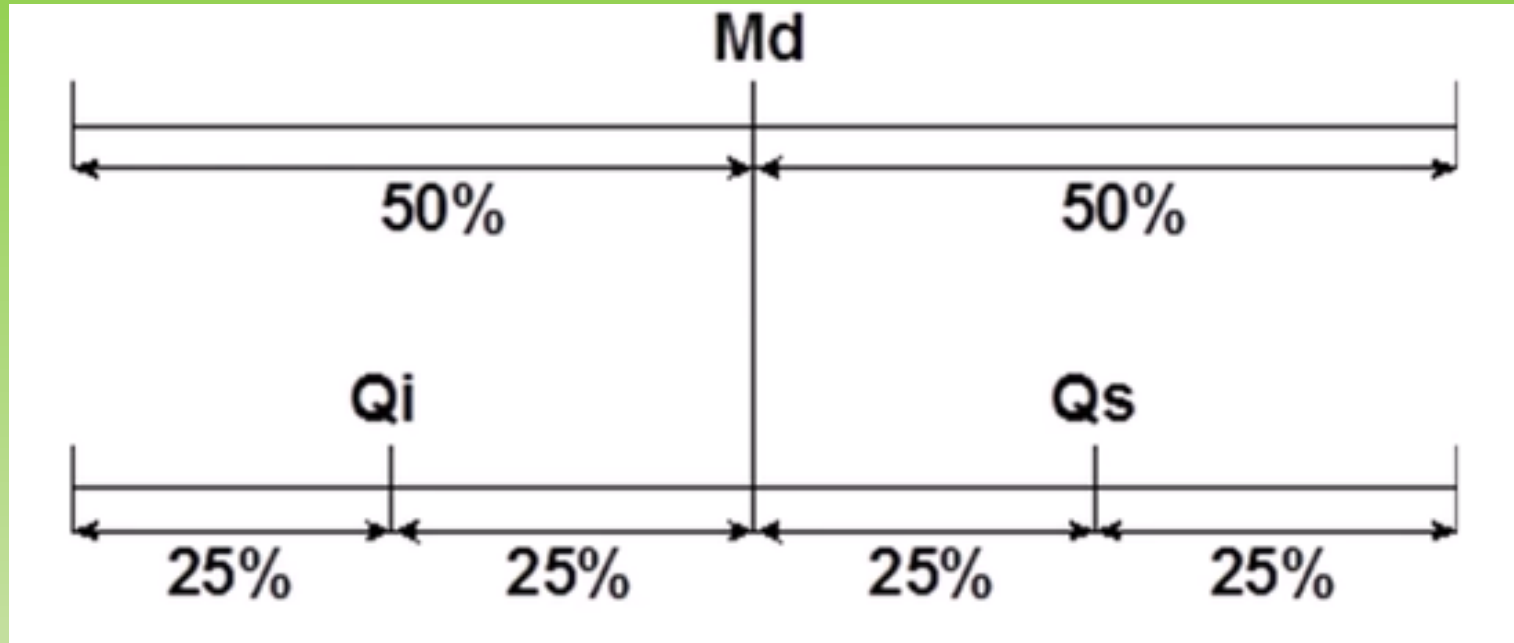
Medidas descritivas:

- Medidas de posição (Média, Mediana, Quartis e Percentis, Moda)
- Medidas de dispersão (Variância, Desvio Padrão, Coeficiente de Variação)

Média aritmética simples \bar{x}

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

Outras medidas de posição – Separatrizes Q_i , Md , Q_s



$Q_2 = Md = \text{mediana}$

$Q_1 = Q_i = 1^\circ. \text{ quartil ou quartil inferior}$

$Q_3 = Q_s = 3^\circ. \text{ quartil ou quartil superior}$

Posição das separatrizes para um conjunto com n valores

Posição da mediana (M_d):

$$\frac{n + 1}{2}$$

Posição do quartil inferior (Q_i)

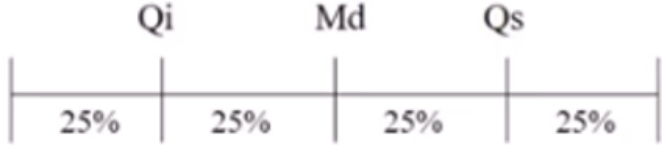
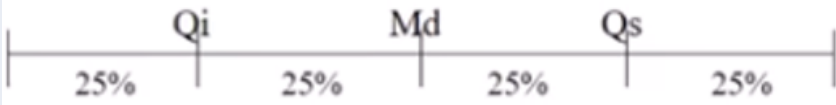

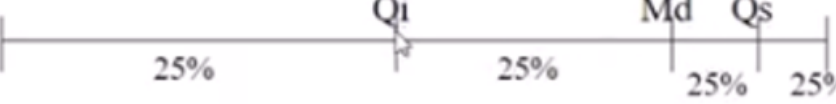
$$\frac{n + 1}{4}$$

Posição do quartil superior (Q_s)

$$\frac{3(n + 1)}{4}$$

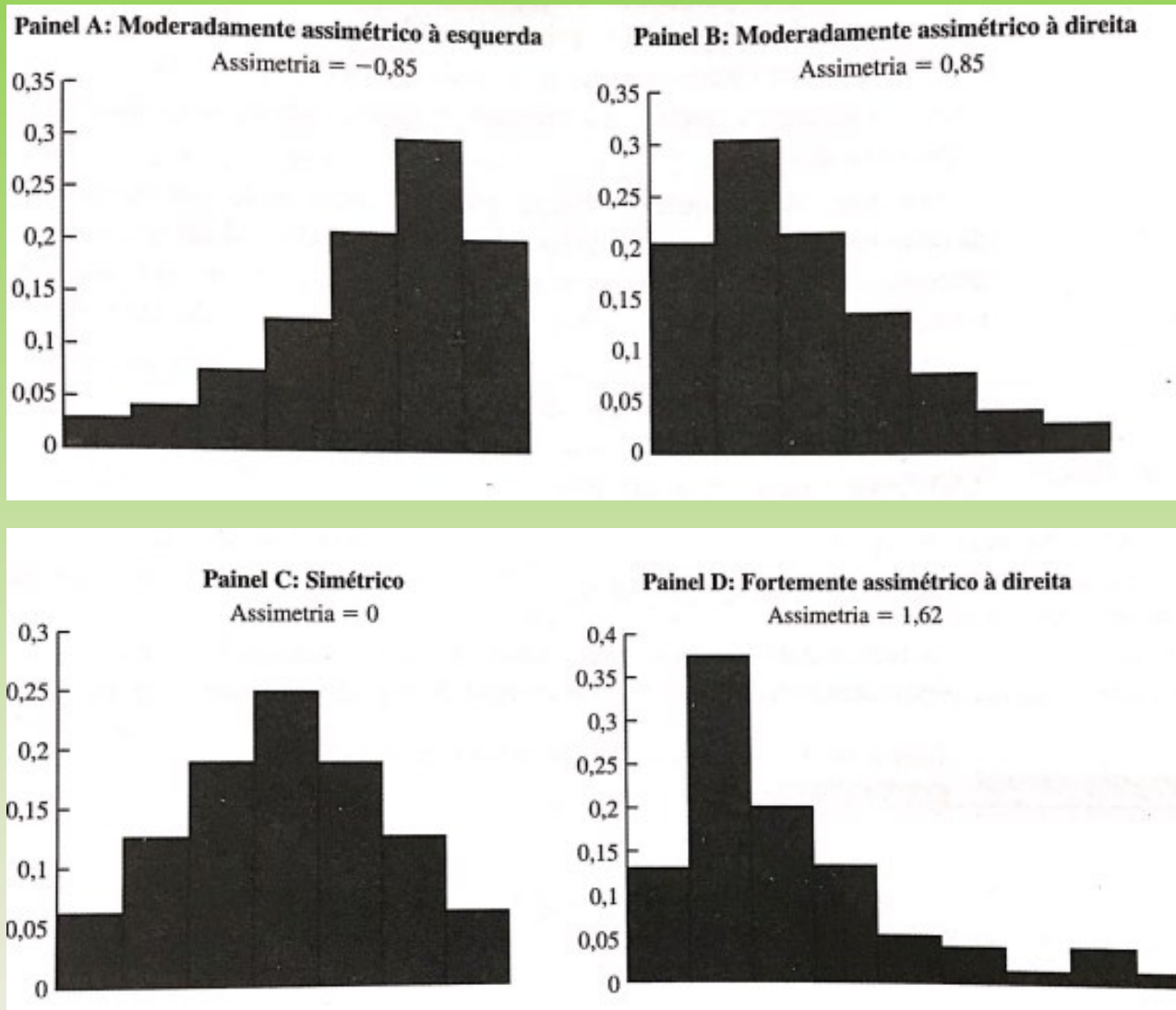
Medidas de posição – Separatrizes Q_i , M_d , Q_s

- Servem para avaliar a assimetria e dispersão dos valores

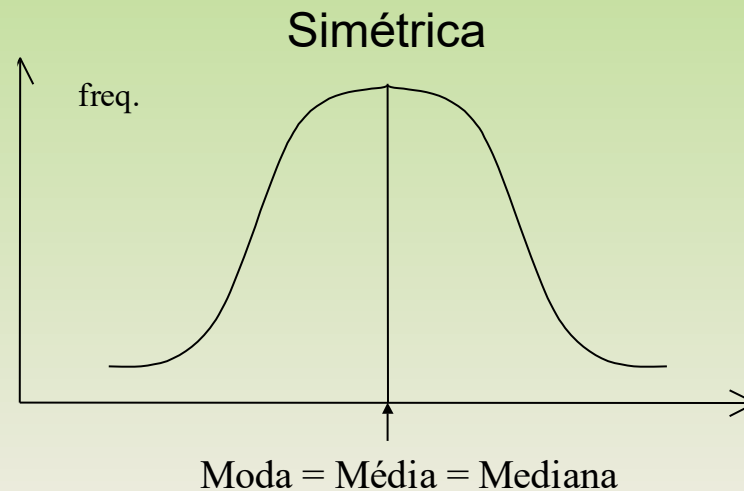
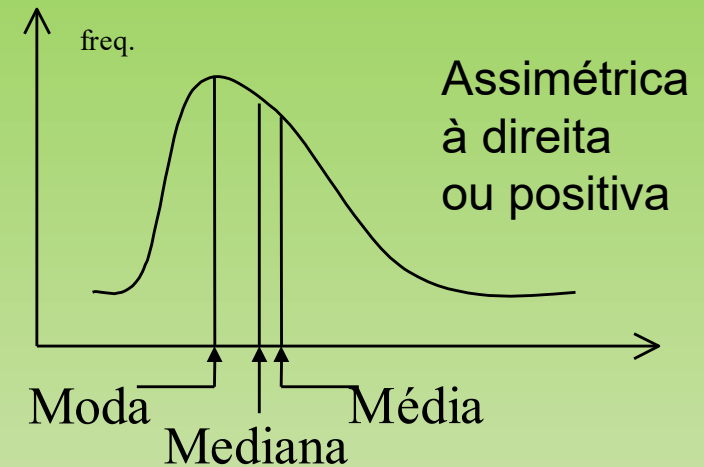
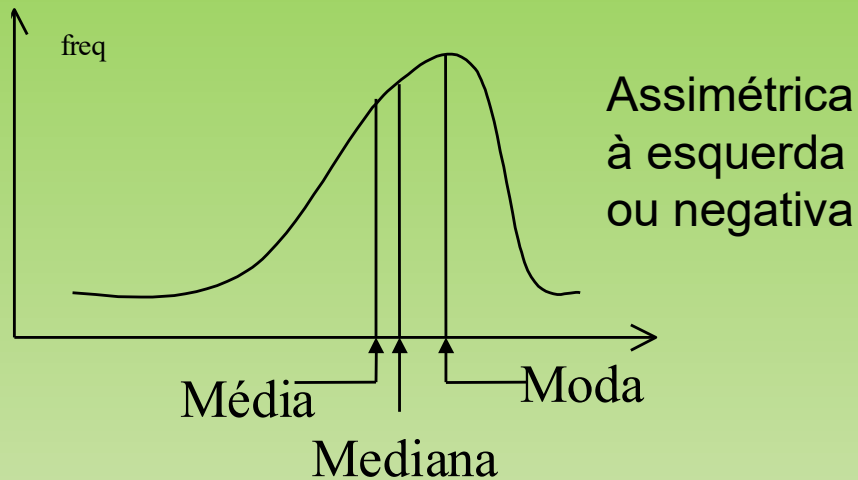
Simétrico	 <p>A horizontal line with four vertical tick marks. Above the marks are labels Q_i, M_d, and Q_s. Below the line, the intervals between the marks are each labeled 25%.</p>
Simétrico, com maior dispersão	 <p>A horizontal line with four vertical tick marks. Above the marks are labels Q_i, M_d, and Q_s. Below the line, the intervals between the marks are each labeled 25%.</p>
Assimétrico à direita	 <p>A horizontal line with four vertical tick marks. Above the marks are labels Q_i, M_d, and Q_s. Below the line, the intervals are labeled 25%, 25%, and 25% from left to right. The distance between Q_i and M_d is the smallest, and the distance between M_d and Q_s is the largest.</p>
Assimétrico à esquerda	 <p>A horizontal line with four vertical tick marks. Above the marks are labels Q_i, M_d, and Q_s. Below the line, the intervals are labeled 25%, 25%, and 25% from left to right. The distance between Q_i and M_d is the largest, and the distance between M_d and Q_s is the smallest.</p>

Medidas de posição – Separatrizes Q_i , M_d , Q_s

- Avaliar a assimetria e dispersão dos valores

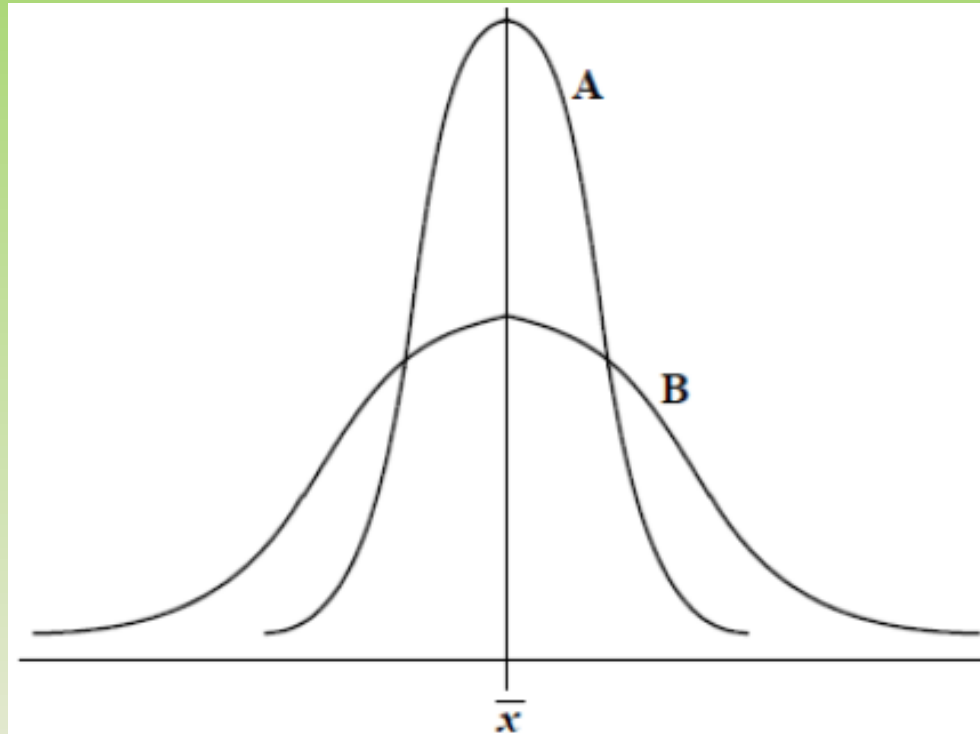


Avaliação da assimetria por média mediana e moda



Medidas de dispersão

Uma “Estatística” de dispersão refere-se à variabilidade ou heterogeneidade dos dados.



Medidas de dispersão para uma amostra de tamanho n

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

Desvio padrão $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

Coeficiente de variação $CV = \frac{s}{\bar{x}} \times 100$

Exemplo 1: Notas da P1 de 40 alunos da TURMA A

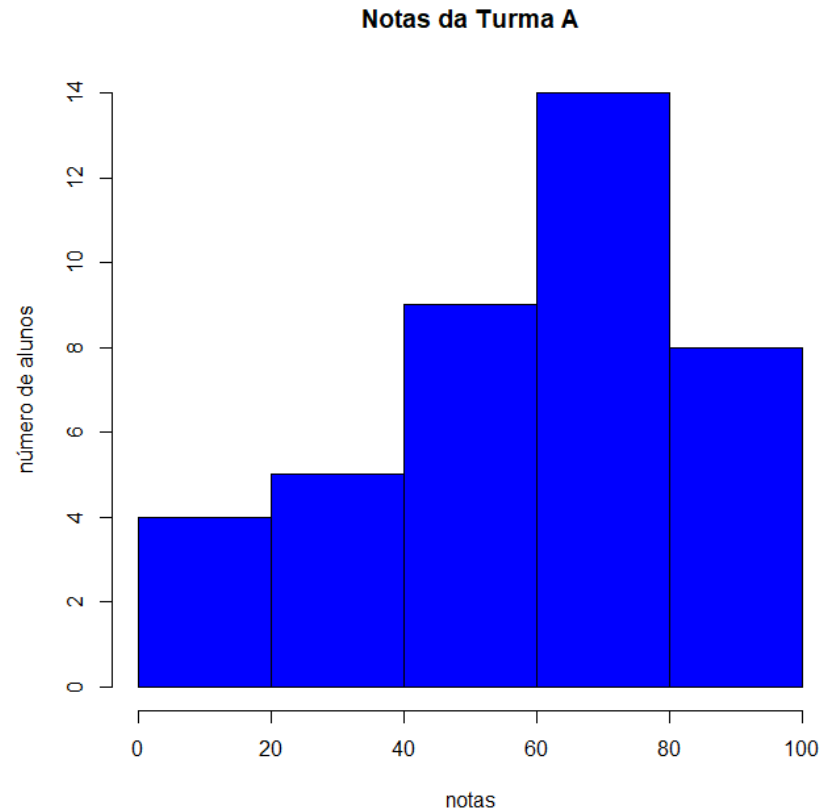
0 10 10 20 30 40 40 40 40 50
50 50 50 50 60 60 60 60 70 70
70 70 70 70 80 80 80 80 80 80
80 80 90 90 90 90 100 100 100 100



Notas	Número de alunos
0	1
10	2
20	1
30	1
40	4
50	5
60	4
70	6
80	8
90	4
100	4
Total	40

14

Exemplo 1:
Notas da P1
de 40 alunos
da TURMA A



Exemplo 1: Notas da P1 de 40 alunos da TURMA A

0 10 10 20 30 40 40 40 40 50
50 50 50 50 60 60 60 60 70 70
70 70 70 70 80 80 80 80 80 80
80 80 90 90 90 90 100 100 100 100



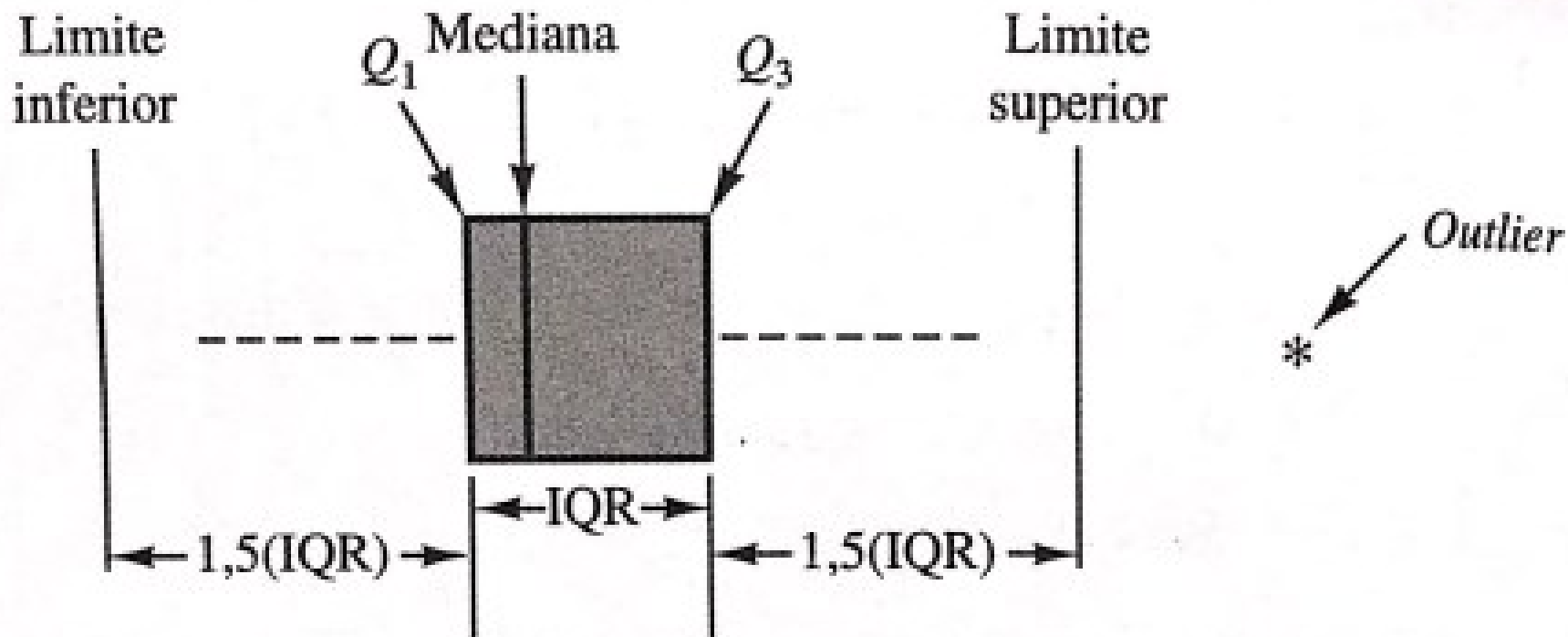
Notas	Número de alunos
0	1
10	2
20	1
30	1
40	4
50	5
60	4
70	6
80	8
90	4
100	4
Total	40

a) Calcular as medidas descritivas:

Média, Mediana, Moda, Q1, Q3, desvio padrão e coeficiente de variação

b) Construa o *Box Plot* (Diagrama em caixas) das notas e classifique a distribuição dos dados em relação à assimetria.

Elementos do Box Plot



FONTE: SWEENEY, D.J; WILLIAMS, T.A.; ANDERSON, D.R. Estatística Aplicada à Administração e economia. 6ed. Cengage, 2013.

$$IQR = Q_S - Q_I$$

*IQR ou DIQ = amplitude interquartil
distância interquartil
desvio interquartilico*

Exemplo 2: Notas da P1 de 70 alunos da TURMA B

5	5	5	5	5	5	5	5	5	8
10	10	10	10	10	15	15	20	20	20
20	20	20	20	25	25	25	25	25	30
30	30	35	35	35	35	35	35	35	40
40	40	40	40	45	45	45	45	50	50
50	50	50	50	55	55	55	60	60	70
70	80	85	90	90	95	100	100	100	100

Realize uma análise exploratória das notas, verifique se há valores discrepantes (outliers) e discuta a assimetria dos dados.



Exemplo 2: Notas da P1 de 70 alunos da TURMA B



Sintaxe no R para o exemplo 2

```
TURMA_B <-  
c(rep(5,9),8,rep(10,5),rep(15,2),rep(20,7),rep(25,5),rep(30,3),rep(35,7),rep(40,5)  
,rep(45,4),rep(50,6),rep(55,3),rep(60,2),rep(70,2),80,85,rep(90,2),95,rep(100,4))
```

Exemplo 3: Fazer o comparativo das notas para das turmas A e B. Utilize o box plot



sintaxe está no arquivo `sintaxe_R_aula6`

Dados no SIGA-A

Assimetria

Quando média e mediana são diferentes: há assimetria.

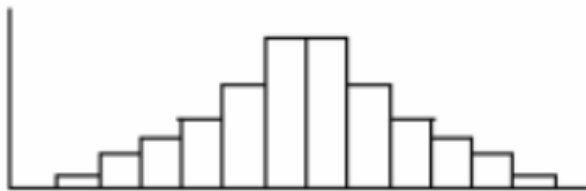
Medida de assimetria:

$$\text{Assimetria} = \frac{\mathbf{n} \times \sum_{i=1}^{\mathbf{n}} (\mathbf{x}_i - \bar{\mathbf{x}})^3}{[(\mathbf{n} - 1) \times (\mathbf{n} - 2) \times \mathbf{s}^3]}$$

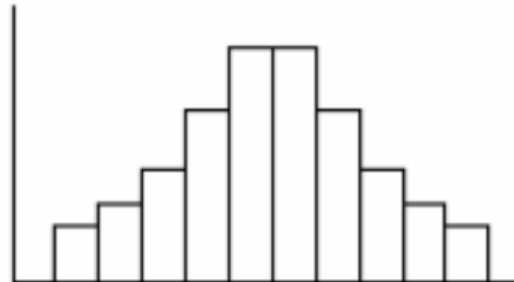
- Se assimetria = 0, a distribuição é SIMÉTRICA.
- Assimetria > 0, a distribuição é assimétrica positiva ou à direita.
- Assimetria < 0, a distribuição é assimétrica negativa ou à esquerda

Curtose

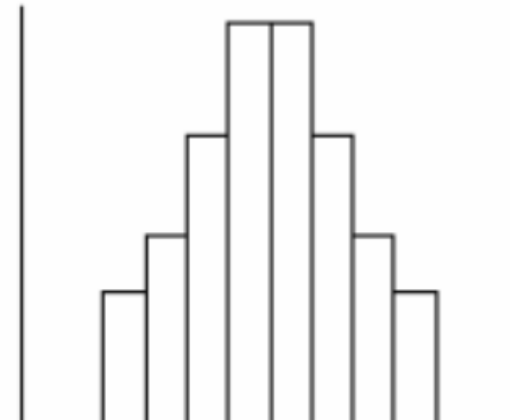
Medida do “achatamento” da distribuição:



Platicúrtica



Mesocúrtica



Leptocúrtica

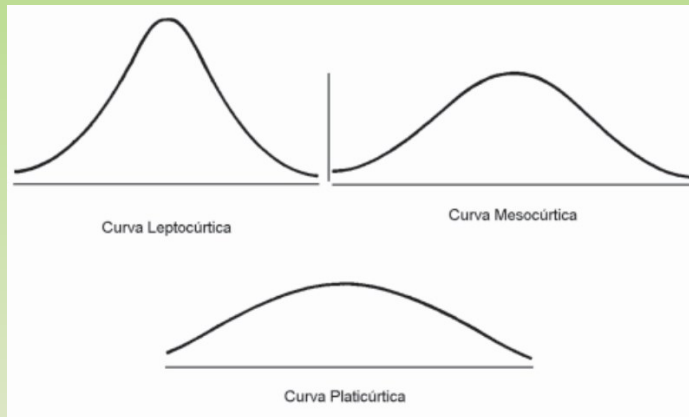
Curtose (K)

Medida do “achatamento” da distribuição:

Mesocúrtica: achatamento equivalente ao da **curva normal**, curtose = 0.

Leptocúrtica: curva afilada, com pico elevado, curtose > 0.

Platicúrtica: curva bem achatada, curtose < 0.



$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$



$K = 0,263$ – Distribuição Mesocúrtica

$K > 0,263$ – Distribuição Platicúrtica

$K < 0,263$ – Distribuição Leptocúrtica

Onde:

K = Coeficiente de curtose

Q_1 = 1º quartil

Q_3 = 3º quartil

P_{90} = Percentil 90

P_{10} = Percentil 10