

**UNIVERSIDADE FEDERAL DO PARÁ  
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

**Tópicos Especiais em Computação: Aprendizado  
de Máquina**

**Cap 2: Análise e Pré-processamento de dados**

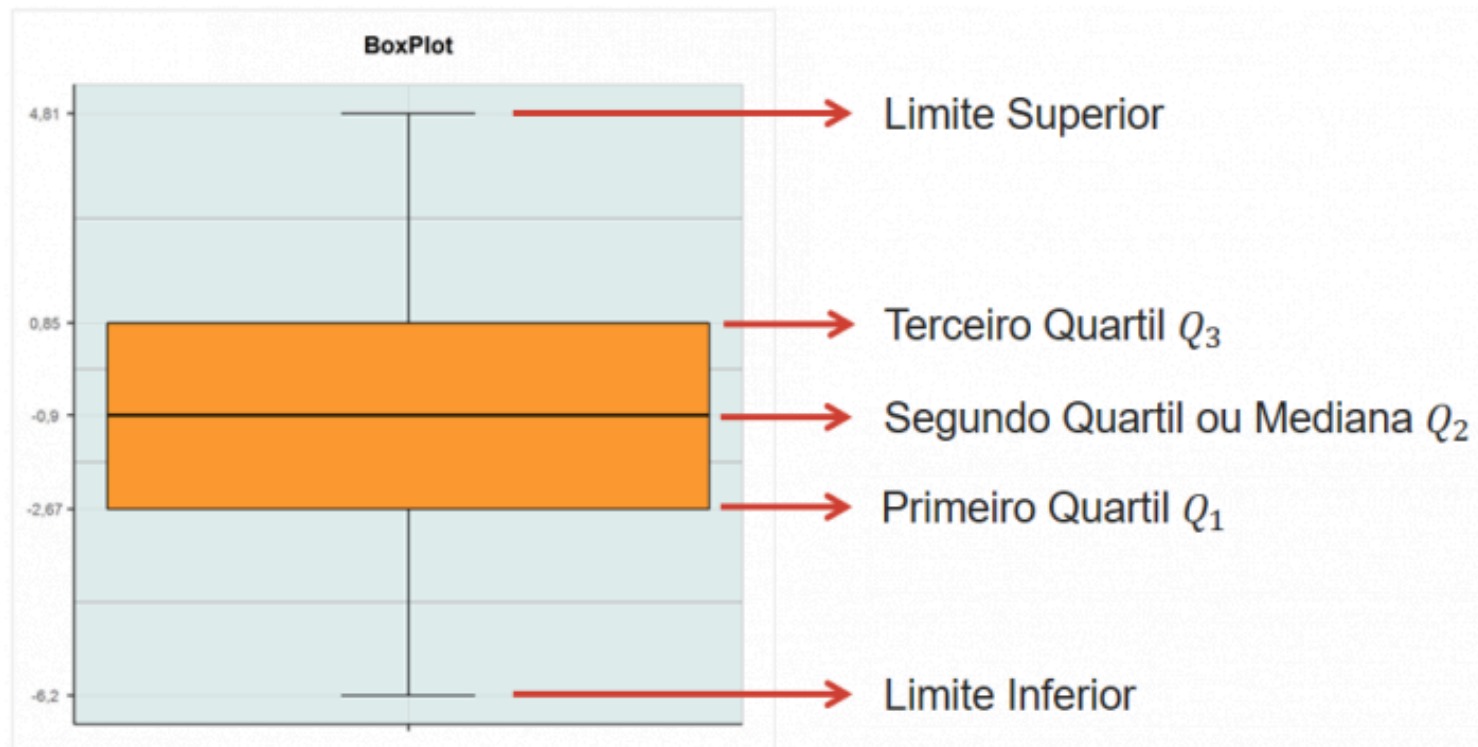
**Prof. Jefferson Moraes  
Email: [jmoraes@ufpa.br](mailto:jmoraes@ufpa.br)**

# Boxplots

- Também chamado de diagramas de Box e Whisker
- formado pelo primeiro e terceiro quartil e pela mediana

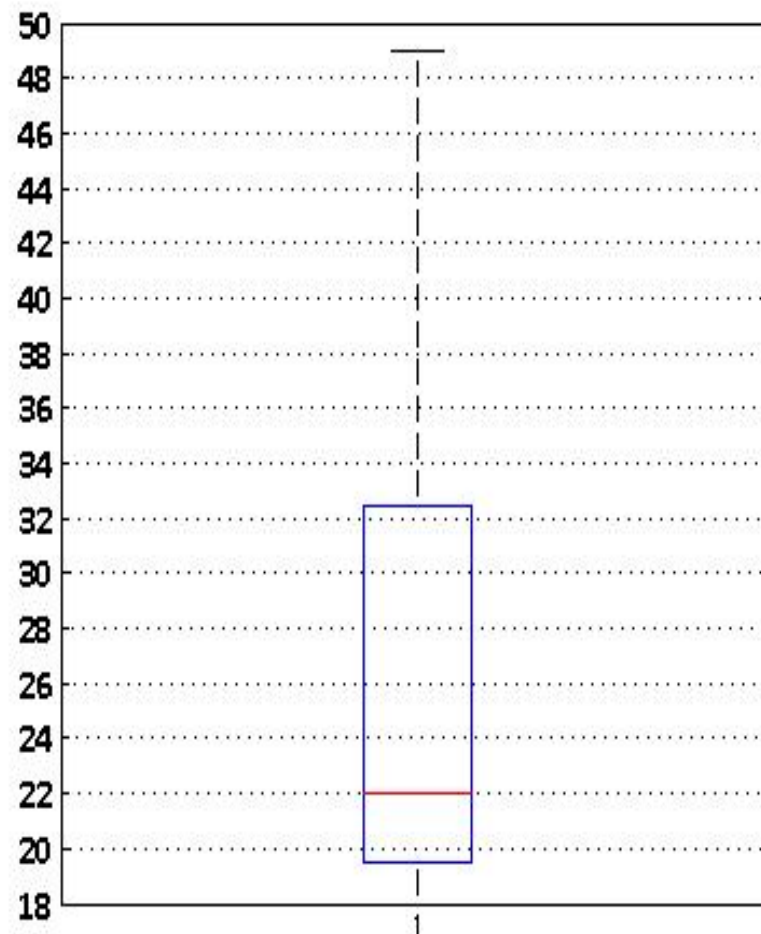
Limite inferior:  $\max\{\min(\text{dados}); Q_1 - 1,5(Q_3 - Q_1)\}$ .

Limite superior:  $\min\{\max(\text{dados}); Q_3 + 1,5(Q_3 - Q_1)\}$ .



# Boxplot

Id.	Nome	Idade
4201	João	28
3217	Maria	18
4039	Luiz	49
1920	José	18
4340	Cláudia	21
2301	Ana	22
1322	Marta	19
3027	Paulo	34



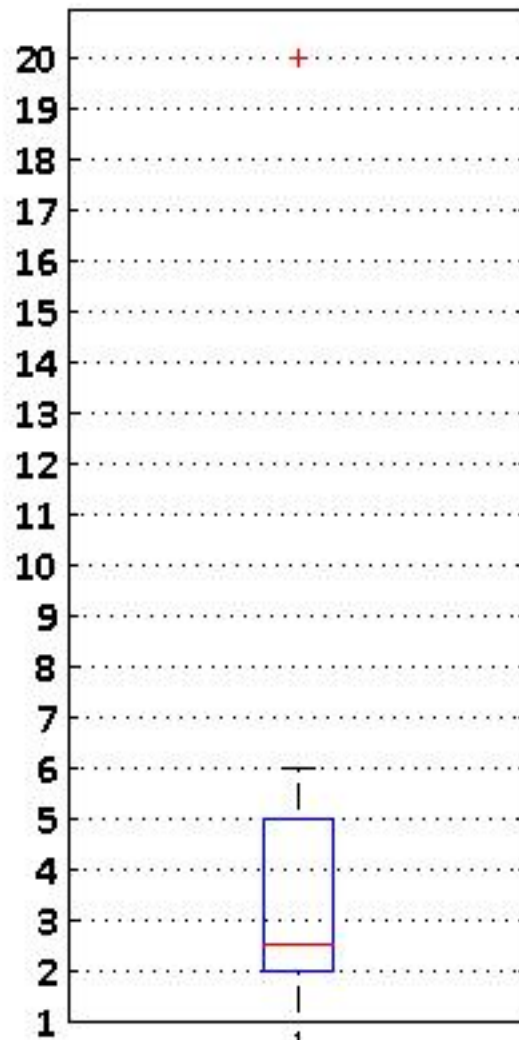
Est. Diagnóstico	
SP	Doente
MG	Doente
RS	Saudável
MG	Doente
PE	Saudável
RJ	Doente
AM	Doente
GO	Saudável

# Boxplot

*Outlier*



Id.	Nome	Idade	Sexo
4201	João	28	M
3217	Maria	18	F
4039	Luiz	49	M
1920	José	18	M
4340	Cláudia	21	F
2301	Ana	22	F
1322	Marta	19	F
3027	Paulo	34	M

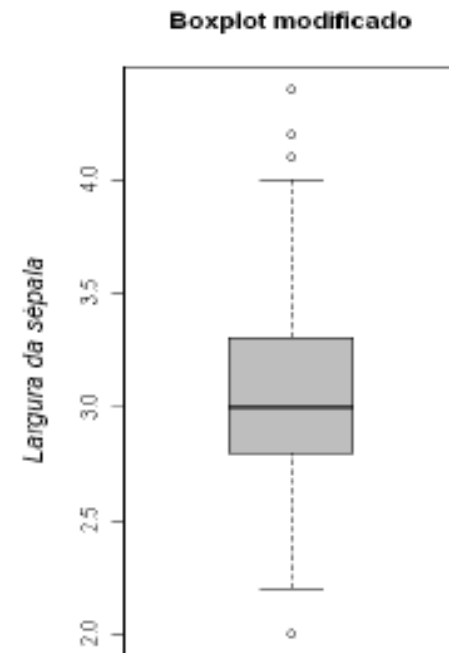
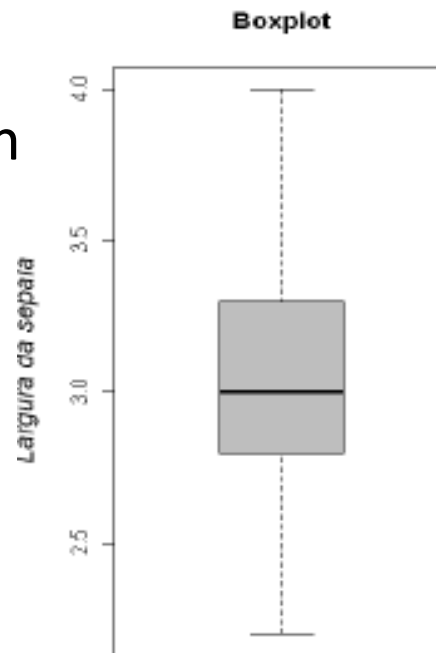


# Int.	Est.	Diagnóstico
2	SP	Doente
4	MG	Doente
2	RS	Saudável
20	MG	Doente
1	PE	Saudável
3	RJ	Doente
6	AM	Doente
2	GO	Saudável

# Boxplot

- Ex: conjunto de dados iris

- 150 instâncias
- 4 atributos de entrada (con
  - Tamanho pétala
  - Tamanho sépala
  - Largura pétala
  - Largura sépala
- 3 classes (espécies de íris)
  - Íris vírginica
  - Íris setosa
  - Íris versicolor

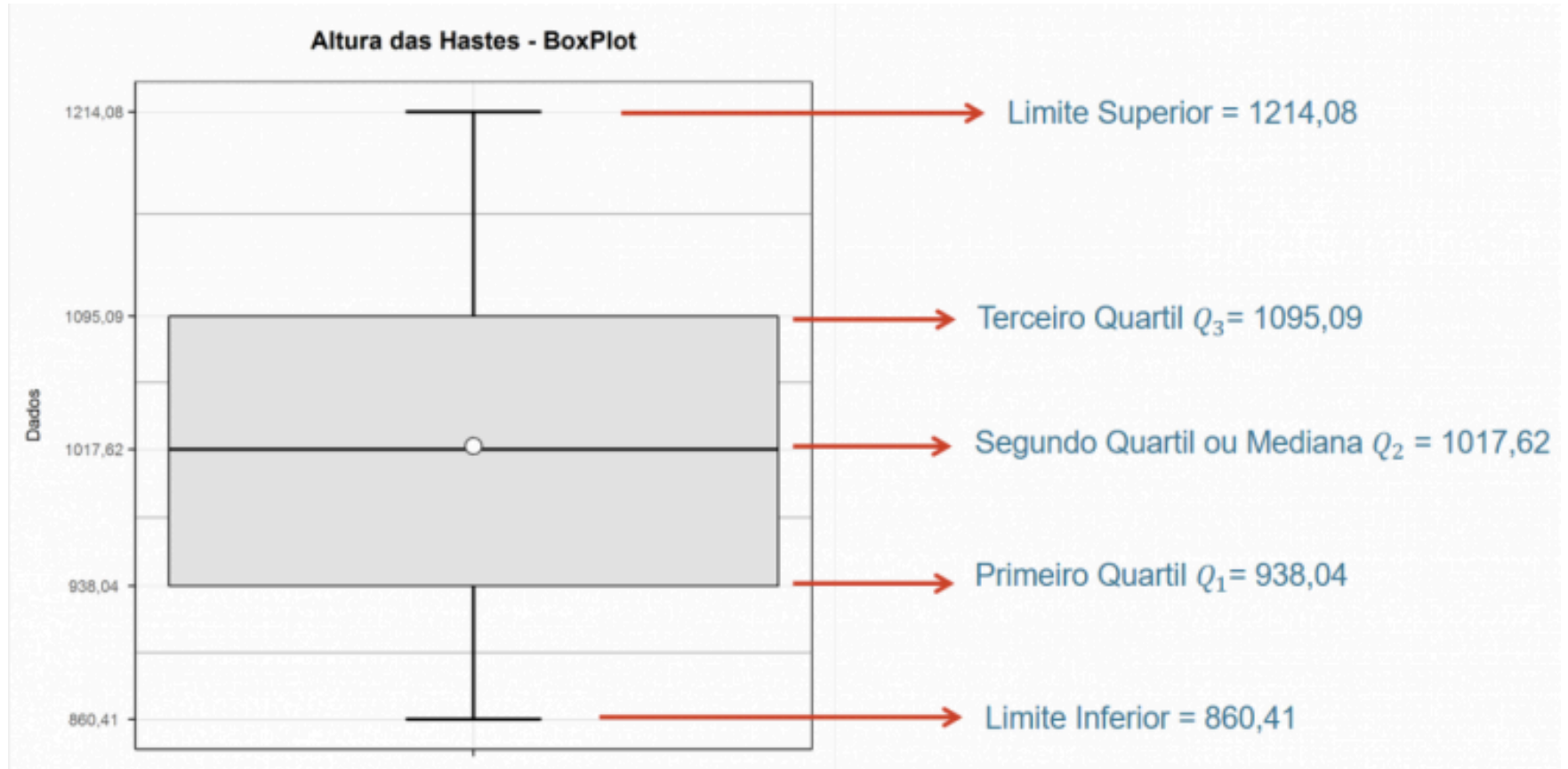


# Exercicio de fixação

- Na Tabela a seguir temos as medidas da altura de 20 hastes. Faça o boxplot correspondente.

Dados da usinagem			
903,88	1036,92	1098,04	1011,26
1020,70	915,38	1014,53	1097,79
934,52	1214,08	993,45	1120,19
860,41	1039,19	950,38	941,83
936,78	1086,98	1144,94	1066,12

# Exercício de fixação (resposta)



# Dados univariados: medidas de espalhamento

- Medem dispersão ou espalhamento de um conjunto de valores
  - Permitem observar se valores estão
    - Espalhados (dispersos)
    - Concentrados em torno de um valor (ex. Média)
- Medidas comuns
  - Intervalo
  - Variância
  - Desvio padrão



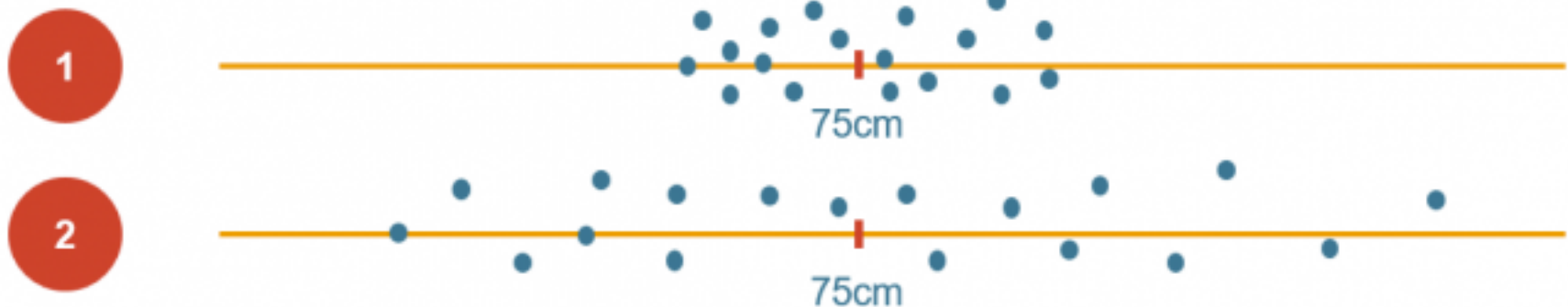
# Intervalo

- Mostra a dispersão máxima entre os valores
  - Medida simples
- Problema: não é uma medida boa se a maioria dos valores estão próximos de um ponto, com um pequeno número de valores extremos

$$\text{intervalo}(\mathbf{x}) = \max_{i=1, \dots, n}(x_i) - \min_{i=1, \dots, n}(x_i)$$

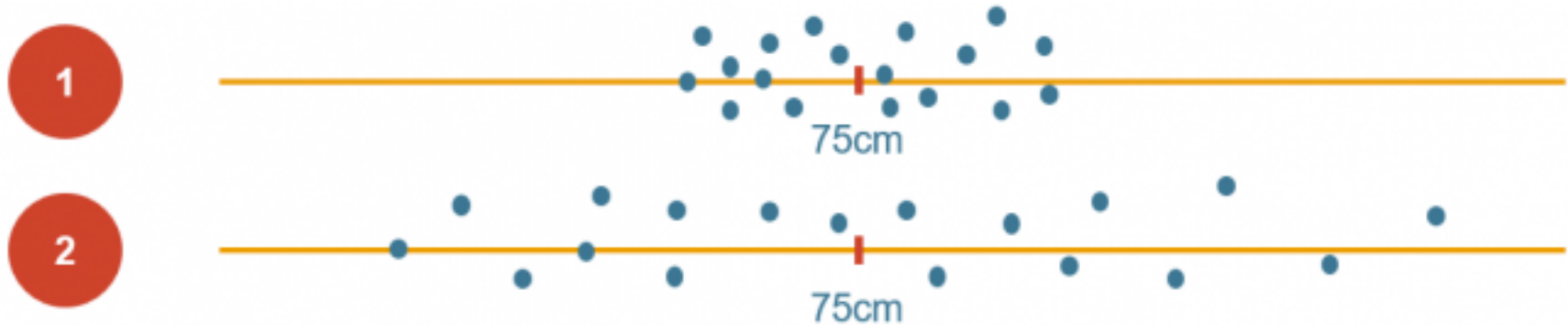
# Medidas de dispersão

- Considere o exemplo de duas linhas de produção de uma peça
- A medida média do comprimento da peça é de 75cm e ambas as linhas estão produzindo peças com médias próximas desse valor
- Podemos considerar que as peças produzidas por ambas as linhas são adequadas?



# Medidas de dispersão

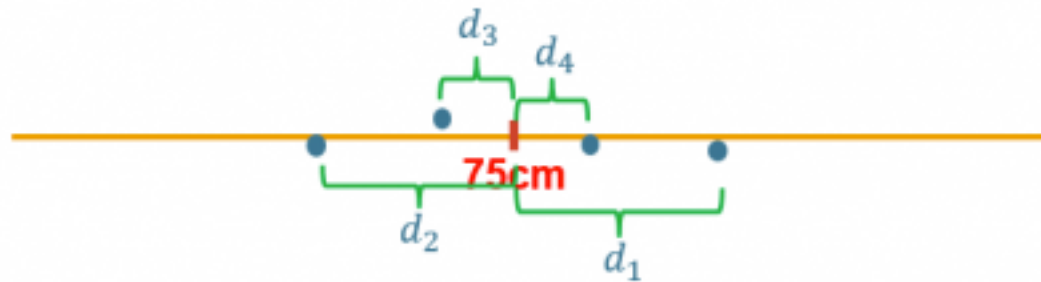
- As peças produzidas pela primeira linha de produção são melhores que a segunda
- Isso ocorre porque a dispersão dos elementos em torno da média é menor, ou seja, os elementos estão mais concentrados em torno da média na primeira linha de produção



# Medidas de dispersão

- Como queremos avaliar a dispersão dos dados em torno da média, esse valor estará relacionado com a distância dos dados em relação à média. Essa distância será chamada de desvio

$$d_i = X_i - \bar{X}$$



- No exemplo, temos.

$$d_1 + d_2 + d_3 + d_4 = 0$$

- O qual nos levaria à conclusão errada de que não existe variação entre os dados.
- Dispersão é sinônimo de variação ou variabilidade. Para medir a dispersão, duas medidas são usadas mais frequentemente: a **amplitude** e o **desvio padrão**.

# Amplitude

- É definida como sendo a diferença entre o maior e o menor valor do conjunto de dados
- Denotaremos a amplitude por R ou H
- Portanto, consideremos o conjunto de dados ordenado

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$$

- A amplitude R dos dados é dada por:

$$R = X_{(n)} - X_{(1)}$$

# Desvio padrão

- Para definirmos desvio padrão é necessário definir variância. A notação mais comumente usada é:

$s^2$  - variância amostral.

$\sigma^2$  - variância populacional.

$s$  - desvio padrão amostral.

$\sigma$  - desvio padrão populacional.

# Variância populacional

- A variância de uma população  $\{x_1, \dots, x_N\}$  de  $N$  elementos é a medida de dispersão definida como a média do quadrado dos desvios dos elementos em relação à média populacional  $\mu$ . Ou seja, a variância populacional é dada por:

$$\sigma^2 = \sum_{i=1}^N \frac{(x_i - \mu)^2}{N}$$

# Variância amostral

- A variância de uma amostra  $\{x_1, \dots, x_n\}$  de  $n$  elementos é definida como a soma ao quadrado dos desvios dos elementos em relação à sua média dividido por  $n-1$

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

- Por que dividido por  $n-1$  (vide <https://www.ime.usp.br/~belitsky/wiki/lib/exe/fetch.php?media=mae121:denominadornoestimadordavariancia.pdf>)



# Variância e desvio padrão

- Mais utilizados      Valor esperado

$$E[\mathbf{X}] = \mu = \sum_{i=1}^n p_i x_i$$

$$\text{Var}(\mathbf{X}) = \sigma^2 = E[(\mathbf{X} - \mu)^2]$$

$$\text{Desvio padrão}(\mathbf{X}) = \sigma = \sqrt{\sigma^2}$$

- Problema: são distorcidas pela presença de outliers

Prove que:

$$\text{Var}(\mathbf{X}) = E[X^2] - E[X]^2$$

# Desvio padrão

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

**Desvio\_padrão= 10,8**

**Desvio\_padrão = 6,3**

# Outras medidas

- Desvio médio absoluto (*Average absolute deviation*)

$$AAD(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|$$

- Desvio mediano absoluto (*Median absolute deviation*)

$$MAD(\mathbf{X}) = \text{mediana}(\{|x_1 - \mu|, \dots, |x_n - \mu|\})$$

- Intervalo interquartil (*interquartil range*)

$$IQR(\mathbf{X}) = P_{75\%} - P_{25\%}$$

# Outras medidas de espalhamento

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Intervalo = 31  
Desvio\_padrao = 10,8  
AAD = 8,2  
MAD = 3,5  
IQR = 14,3

Intervalo = 19  
Desvio\_padrao = 6,3  
AAD = 4  
MAD = 1  
IQR = 3,5

# Momento

- Medidas em torno da média (média e desvio padrão), são em sua maioria instâncias de medida de momento

$$\text{momento}_k(\mathbf{X}) = \frac{\sum_{i=1}^n (x_i - \mu)^k}{(n-1)}$$

- Para cada valor de  $k$ , uma medida diferente de momento é definida
  - $k=1 \rightarrow$  **momento central** (primeiro momento em torno da origem)
  - $K=2 \rightarrow$  **variância** (segundo momento central)
  - $K=3 \rightarrow$  **obliquidade** (terceiro momento central)
  - $K=4 \rightarrow$  **curtose**(quarto momento central)

# Obliquidade e curtose

- São medidas de distribuição, por mostrarem como os valores estão distribuídos
  - Obliquidade ou assimetria(*skewness*)
    - Mede a simetria da distribuição em torno da média
  - Curtose (Kurtosis)
    - Captura o achatamento da curva da função de distribuição de probabilidade

# Representação do conjunto de dados

- Distribuições de frequência

- ▶ Frequência relativa
- ▶ Frequência acumulada

- Representação Gráfica

- ▶ Histogramas

# Organização de dados

- Os métodos utilizados para organizar dados compreendem o arranjo desses dados em subconjuntos que apresentem características similares
  - mesma idade (ou “faixa etária”), mesma finalidade, mesma escola, mesmo bairro, etc
- Os *dados agrupados* podem ser resumidos em tabelas ou gráficos e, a partir desses, podemos obter as estatísticas descritivas já definidas: média, mediana, desvio, etc.
- Dados organizados em grupos ou categorias/classes são usualmente designados “*distribuição de freqüência*”.





# Distribuição de frequência

- Uma distribuição de frequência é um método de se agrupar dados em classes de modo a fornecer a quantidade (e/ou a percentagem) de dados em cada classe
- Com isso, podemos *resumir e visualizar* um conjunto de dados sem precisar levar em conta os valores individuais
- Uma *distribuição de frequência (absoluta ou relativa)* pode ser apresentada em tabelas ou gráficos

# Construindo uma distribuição de frequência

- Adotemos o conjunto de dados que represente a população
- Ordene em ordem crescente ou decrescente

Eventos	Altura
Aluno 1	1,60
Aluno 2	1,69
Aluno 3	1,72
Aluno 4	1,73
Aluno 5	1,73
Aluno 6	1,74
Aluno 7	1,75
Aluno 8	1,75
Aluno 9	1,75
Aluno 10	1,75
Aluno 11	1,75
Aluno 12	1,76
Aluno 13	1,78
Aluno 14	1,80
Aluno 15	1,82
Aluno 16	1,82
Aluno 17	1,84
Aluno 18	1,88

# Construindo uma distribuição de frequência

- Determine a quantidade de classes (k)
  - Regra de Sturges (Regra do logaritmo)
    - $k = 1 + 3,3 \log(n)$
  - Regra da potência de 2
    - $k =$  menor valor inteiro tal que  $2^k \geq n$
  - Regra da raiz quadrada
    - $K = \sqrt{n}$
  - Bom senso!!
    - Decida a quantidade de classes que garanta observar como os valores se distribuem

# Construindo uma distribuição de frequência

Regra de Sturges (Logaritmo)		Regra da Potência de 2		Bom Senso		
Quantidade de dados (n)	Quantidade de Classes (k)	Quantidade de dados (n)	Quantidade de Classes (k)	Quantidade de dados (n)	Quantidade MÍNIMA de Classes (k)	Quantidade MÁXIMA de Classes (k)
1	1	1 e 2	1	até 50	5	10
2	2	3 e 4	2	51 a 100	8	16
3 a 5	3	5 a 8	3	101 a 200	10	20
6 a 11	4	9 a 16	4	201 a 300	12	24
12 a 23	5	17 a 32	5	301 a 500	15	30
24 a 46	6	33 a 64	6	mais de 500	20	40
47 a 93	7	65 a 128	7			
94 a 187	8	129 a 256	8			
188 a 376	9	257 a 512	9			
377 a 756	10	513 a 1024	10			

# Construindo uma distribuição de frequência

- Calcule a amplitude das classes (h)
  - Calcule a amplitude do conjunto de dados
    - $L = x_{\max} - x_{\min}$
  - Calcule a amplitude (largura) da classe
    - $h = L/k$
    - Arredonde convenientemente
  - Calcule os limites das classes
    - Classe 1:  $x_{\min}$  até  $x_{\min+h}$
    - Classe 2:  $x_{\min+h}$  até  $x_{\min+2h}$
    - .....
    - Classe k:  $x_{\min+(k-1)h}$  até  $x_{\min+kh}$

# Construindo uma distribuição de frequência

## Limite das classes

- ▶ Utilize a notação:
  - $[x,y)$  – intervalo de entre  $x$  (fechado) até  $y$  (aberto)
- ▶ Freqüentemente temos que “arredondar” a amplitude das classes e, conseqüentemente, arredondar também os limites das classes.
- ▶ Como sugestão, podemos tentar, se possível, um ajuste simétrico nos limites das classes das pontas (i.e., primeira e última) nas quais, *usualmente*, a quantidade de dados é menor.

## Ponto médio das classes

▶ 
$$x_k = L_{\text{inferior}} + (L_{\text{superior}} - L_{\text{inferior}}) / 2$$

# Construindo uma distribuição de freqüência

## Determinação da freqüência das classes

- Consiste em agrupar os dados em cada classe e contar os totais

## Traçar o gráfico

- Dividir o eixo horizontal em tantas partes quanto for o número de classes. *Sugestão: deixe espaço entre o eixo vertical e a primeira classe.*
- Identifique a maior freqüência da classe na tabela e marque esse número (ou outro um pouco maior) na extremidade do eixo vertical; divida esse eixo em algumas partes e marque os valores correspondentes
- Desenhe um retângulo, para cada classe, com largura igual à largura da classe e com altura igual à freqüência da classe

# Exemplo

Do nosso exemplo:

- ▶ Ordenamos os dados
- ▶ Por Sturges, temos:
  - $n=18$  ;  $k=5$  (número de classes)
- ▶ Amplitude de classes
  - Amplitude do conjunto de dados:  $1,88 - 1,60 = 0,28\text{m}$
  - Amplitude de classes:  $0,28/5 = 0,056$
  - Arredondado  $h = 0,06\text{m}$

Altura
1,60
1,69
1,72
1,73
1,73
1,74
1,75
1,75
1,75
1,75
1,75
1,75
1,76
1,78
1,80
1,82
1,82
1,84
1,88



# Construindo uma tabela de frequência

- Calcule os Limites de Classe
- Arredonde os Limites de Classe nos extremos
  - ▶  $1,9 - 1,88 = 0,02$
  - ▶ Distribua o excesso:
    - $1,60 - 0,01$ ;  $1,88 + 0,01$
  - ▶ Ajuste todas as classes

Amplitude	0,06
Limites inferiores	Limite superior
1,60	1,66
1,66	1,72
1,72	1,78
1,78	1,84
1,84	1,90

Aqui "sobra"  
0,02m!

Altura
1,60
1,69
1,72
1,73
1,73
1,74
1,75
1,75
1,75
1,75
1,75
1,75
1,76
1,78
1,80
1,82
1,82
1,84
1,88

# Construindo uma tabela de frequência

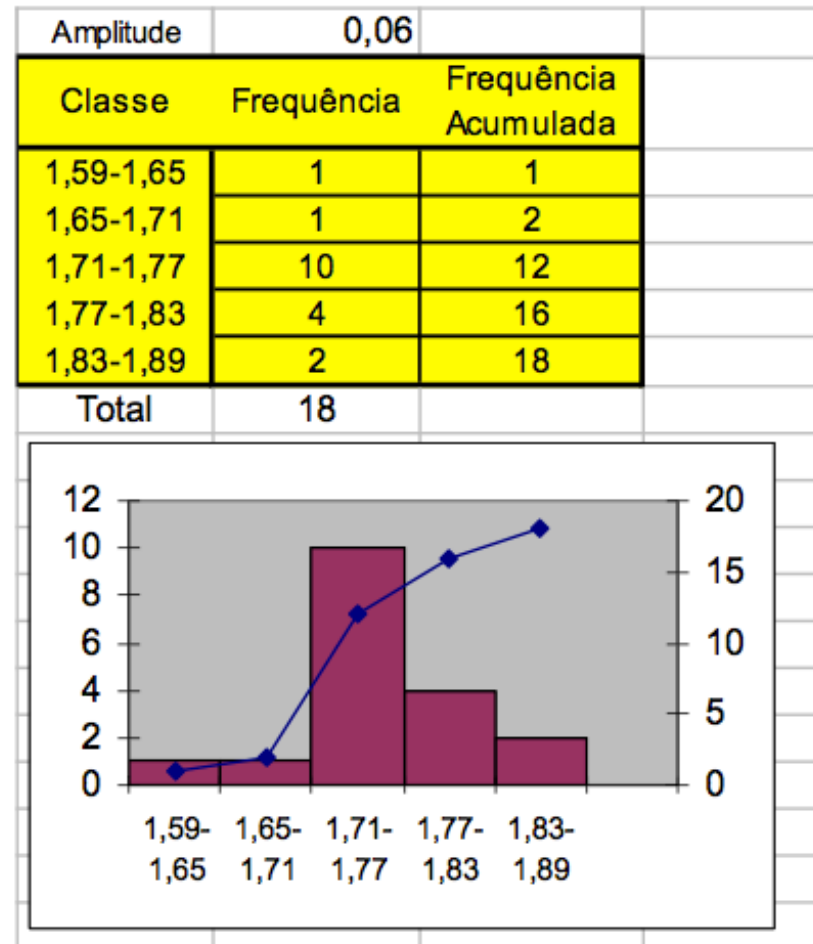
- **Frequências absolutas**
  - ▶ Distribua os eventos ou ocorrência por suas respectivas classes
- **Frequências acumuladas**
  - ▶ Some as ocorrências de dados cumulativamente às classes
- **Observação importante:**
  - ▶ É muito útil representar as frequências em termos percentuais ao total de amostras

Amplitude		0,06	
Dados	Classe	Frequência	Frequência Acumulada
1,60	1,59-1,65	1	1
1,69	1,65-1,71	1	2
1,72	1,71-1,77	10	12
1,73	1,77-1,83	4	16
1,73	1,83-1,89	2	18
1,74	Total	18	
1,75			
1,75			
1,75			
1,75			
1,75			
1,76			
1,78			
1,80			
1,82			
1,82			
1,84			
1,88			

# Representação gráfica

## Histograma

- ▶ Na abscissas, distribua as classes
- ▶ Na ordenada da esquerda, as freqüências absolutas
- ▶ Construa um gráfico de barras para as freqüências
- ▶ Construa um gráfico de linha para a freqüência acumulada (utilize a escala da direita)

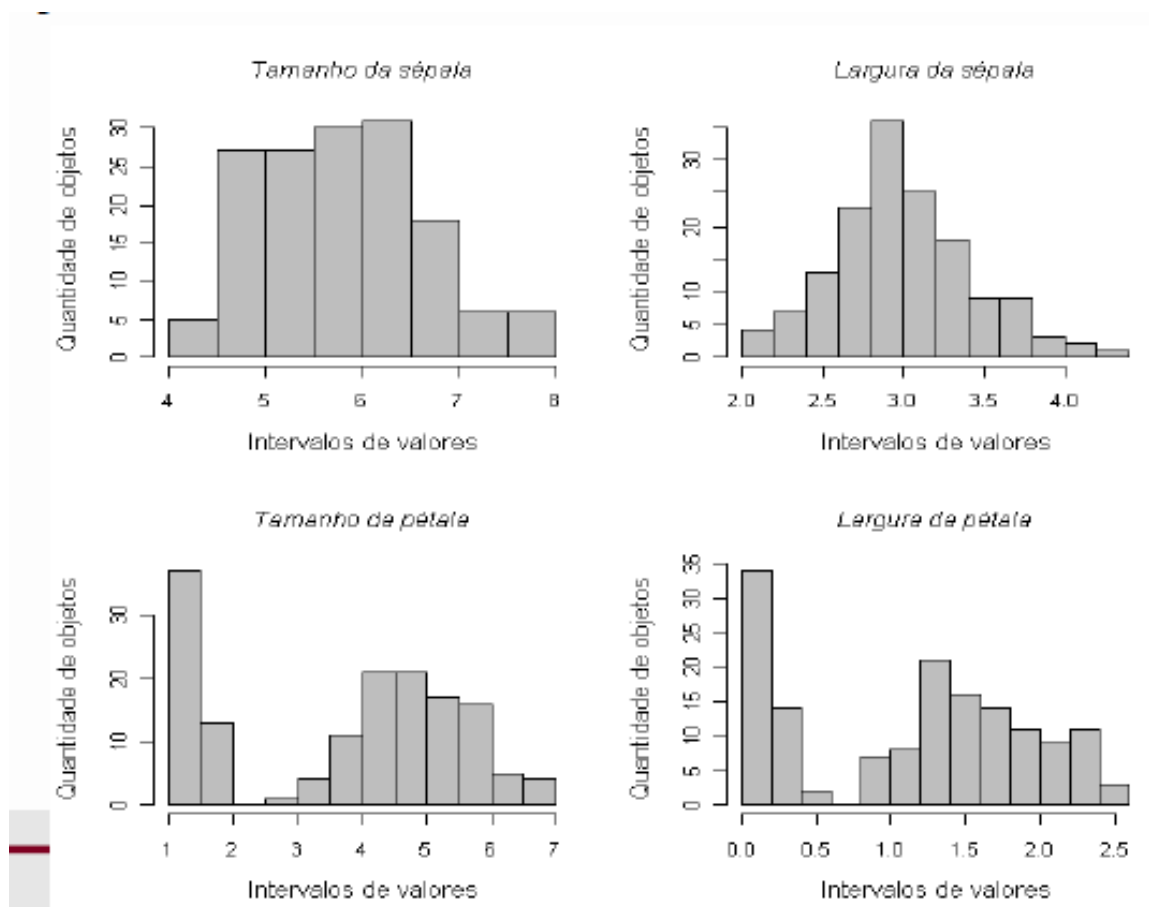


# Histograma

Id.	Nome	Idade
4201	João	28
3217	Maria	18
4039	Luiz	49
1920	José	18
4340	Cláudia	21
2301	Ana	22
1322	Marta	19
3027	Paulo	34

# Histograma

## ■ Base de dados Iris



# Obliquidade

## ■ Equação ( $k=3$ )

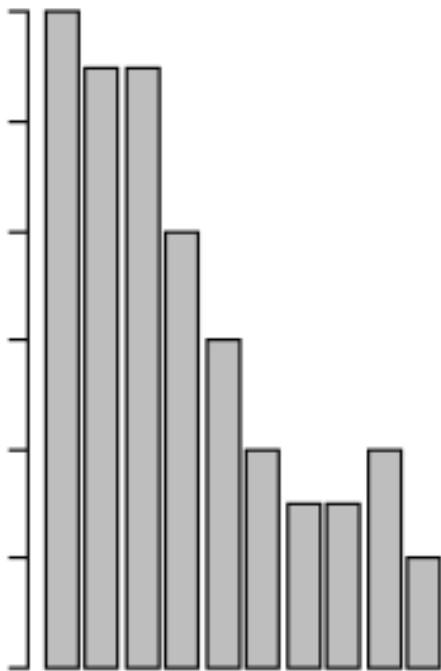
$$\text{obliquidade}(\mathbf{X}) = \text{momento}_3(\mathbf{X}) = \frac{\sum_{i=1}^n (x_i - \mu)^3}{(n-1)\text{desvio\_padrao}^3}$$

## ■ Valores de obliquidade

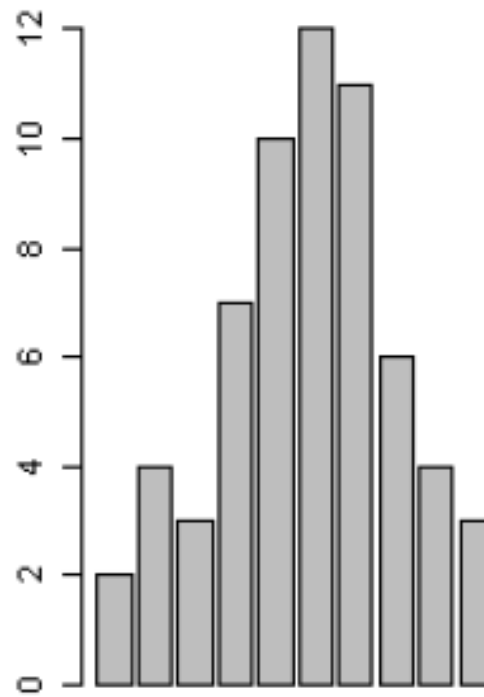
- = 0 (simétrica): distribuição é aproximadamente simétrica
- > 0 (positiva): distribuição concentra-se mais no lado esquerdo
- < 0 (negativa): distribuição concentra-se mais no lado direito

# Obliquidade

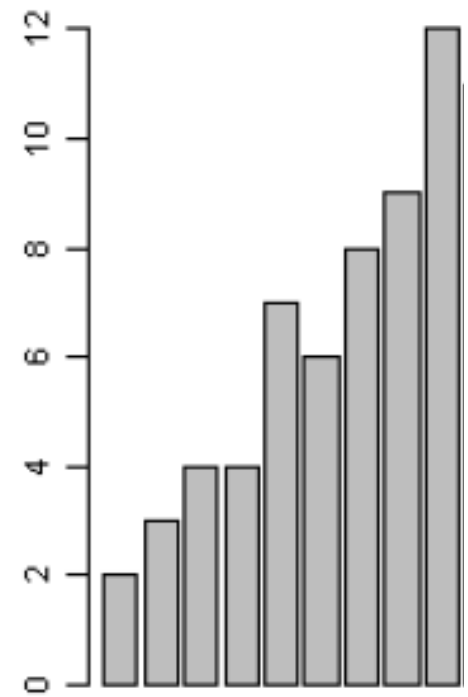
**Positiva**



**Simétrica**



**Negativa**



# Curtose

- Verifica se os dados apresentam um pico ou são achatados em relação a uma distribuição normal

$$\text{curtose}(\mathbf{X}) = \text{momento}_4(\mathbf{X}) = \frac{\sum_{i=1}^n (x_i - \mu)^4}{(n-1)\text{desvio\_padrao}^4}$$

- Para uma distribuição normal com média 0 e variância 1 o valor da curtose é igual a 3. Assim, é feita uma correção na equação

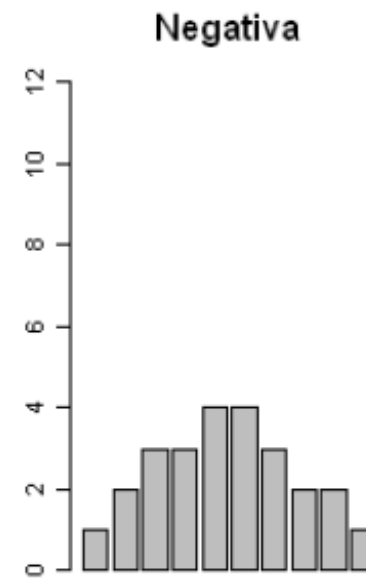
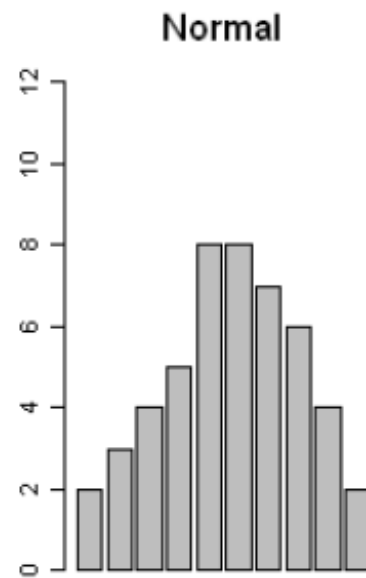
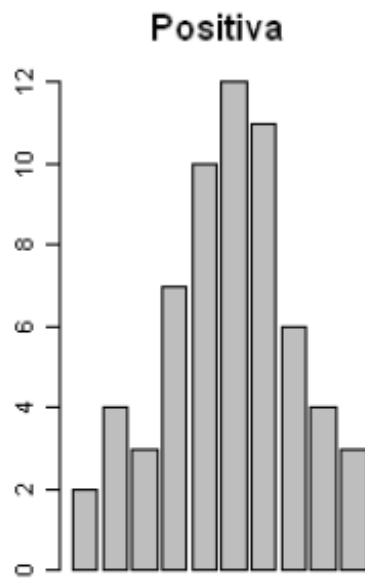
$$\text{curtose}(\mathbf{X}) = \text{momento}_4(\mathbf{X}) = \frac{\sum_{i=1}^n (x_i - \mu)^4}{(n-1)\text{desvio\_padrao}^4} - 3$$



# Curtose

## ■ Valores de curtose

- = 0 (normal): histograma tem achatamento de distribuição normal
- > 0 (positiva): histograma tem distribuição mais alta e concentrada
- < 0 (negativa): histograma tem distribuição mais achatada



# Gráfico de pizza

- Outro gráfico muito utilizado para ilustrar a distribuição de um conjunto de valores
- Indicado para valores quantitativos
  - Para quantitativos, deve agrupar valores em cestas
- Cada valor ocupa fatia com área proporcional ao número de vezes que aparece no conjunto de dados



# Dados multivariados

- São aqueles que possuem mais de uma atributo
  - Ex: conjunto de dados da Iris e do hospital
- Medidas de localidade podem ser obtidas calculando a medida de localidade de cada atributo separadamente
  - Ex: Média para um conjunto de dados com  $d$  atributos

$$\bar{\mathbf{X}} = \left( \bar{x}^1, \dots, \bar{x}^d \right)$$

# Dados multivariados

- Permitem ainda análises da relação entre dois ou mais atributos
  - Para atributos quantitativos, o espalhamento de um conjunto de dados é melhor capturado por uma **matriz de covariância**
    - Cada elemento é a covariância entre dois atributos

# MATRIZ COVARIÂNCIA


A matriz covariância  $\mathbf{K}_{\mathbf{X}}$  associada com um vetor aleatório  $\mathbf{X} = [x_1, x_2, \dots, x_N]^T$  real é expressa como:

$$\mathbf{K}_{\mathbf{X}} = E[(\mathbf{X} - \mathbf{m}_{\mathbf{X}})(\mathbf{X} - \mathbf{m}_{\mathbf{X}})^T]$$

$$\mathbf{K}_{\mathbf{X}} = E \left\{ \begin{bmatrix} (X_1 - m_1) \\ (X_2 - m_2) \\ \dots \\ (X_n - m_n) \end{bmatrix} \begin{bmatrix} (X_1 - m_1) & (X_2 - m_2) & \dots & (X_n - m_n) \end{bmatrix} \right\}$$

$$K_{ij} = E[(X_i - m_i)(X_j - m_j)]$$

$$= E[(X_j - m_j)(X_i - m_i)] = K_{ji}; \quad i, j = 1, \dots, n$$



$$\mathbf{K}_X = E[(\mathbf{X} - \mathbf{m}_X)(\mathbf{X} - \mathbf{m}_X)^T] = E \left\{ \begin{bmatrix} (X_1 - m_1) \\ (X_2 - m_2) \\ \vdots \\ (X_n - m_n) \end{bmatrix} \begin{bmatrix} (X_1 - m_1) & (X_2 - m_2) & \dots & (X_n - m_n) \end{bmatrix} \right\}$$

$$\mathbf{K}_X = \begin{bmatrix} E[(X_1 - m_1)(X_1 - m_1)] & E[(X_1 - m_1)(X_2 - m_2)] & \dots & E[(X_1 - m_1)(X_n - m_n)] \\ E[(X_2 - m_2)(X_1 - m_1)] & E[(X_2 - m_2)(X_2 - m_2)] & \dots & E[(X_2 - m_2)(X_n - m_n)] \\ \dots & \dots & \dots & \dots \\ E[(X_n - m_n)(X_1 - m_1)] & E[(X_n - m_n)(X_2 - m_2)] & \dots & E[(X_n - m_n)(X_n - m_n)] \end{bmatrix}$$

$$\mathbf{K}_X = \begin{bmatrix} E[X_1^2] - m_1^2 & E[X_1 X_2] - m_1 m_2 & \dots & E[X_1 X_n] - m_1 m_n \\ E[X_2 X_1] - m_2 m_1 & E[X_2^2] - m_2^2 & \dots & E[X_2 X_n] - m_2 m_n \\ \dots & \dots & \dots & \dots \\ E[X_n X_1] - m_n m_1 & E[X_n X_2] - m_n m_2 & \dots & E[X_n^2] - m_n^2 \end{bmatrix}$$

$$\mathbf{K}_X = \begin{bmatrix} \sigma_{X_1}^2 & K_{12} & \dots & K_{1n} \\ K_{21} & \sigma_{X_2}^2 & \dots & K_{2n} \\ \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & \sigma_{X_n}^2 \end{bmatrix}$$

# MATRIZ COVARIÂNCIA

- 1 – Se  $X$  é real, todos os elementos de  $K$  são reais.
- 2- Como  $K_{ij}=K_{ji}$ , a matriz covariância pertence à classe das matrizes simétricas.
- 3- Os elementos da diagonal da matriz covariância são as variâncias das variáveis aleatórias, que formam componentes dos vetores.

# MATRIZ CORRELAÇÃO $\mathbf{R}_x$

- Indicação mais clara da força da relação linear entre dois atributos

$$\mathbf{R}_x = E[\mathbf{X}\mathbf{X}^T] = \begin{bmatrix} E[x_1^2] & E[x_1x_2] & \dots & E[x_1x_N] \\ E[x_2x_1] & E[x_2^2] & \dots & E[x_2x_N] \\ \dots & \dots & \dots & \dots \\ E[x_Nx_1] & E[x_Nx_2] & \dots & E[x_N^2] \end{bmatrix}$$



# Covariância e correlação

- Ex: Conjunto de dados iris
- Matriz de covariância

	Tamanho_sépala	Largura_sépala	Tamanho_pétala	Largura_pétala
Tamanho_sépala	0,68569	-0,03927	1,27368	0,51690
Largura_sépala	-0,03927	0,18800	-0,32171	-0,11798
Tamanho_pétala	1,27368	-0,32171	3,11318	1,29639
Largura_pétala	0,51690	-0,11798	1,29639	0,58241

- Matriz de correlação

	Tamanho_sépala	Largura_sépala	Tamanho_pétala	Largura_pétala
Tamanho_sépala	1,00000	-0,10937	0,87175	0,81795
Largura_sépala	-0,10937	1,00000	-0,42052	-0,35654
Tamanho_pétala	0,87175	-0,42052	1,00000	0,96276
Largura_pétala	0,81795	-0,35654	0,96276	1,00000

# Dados multivariados: visualização

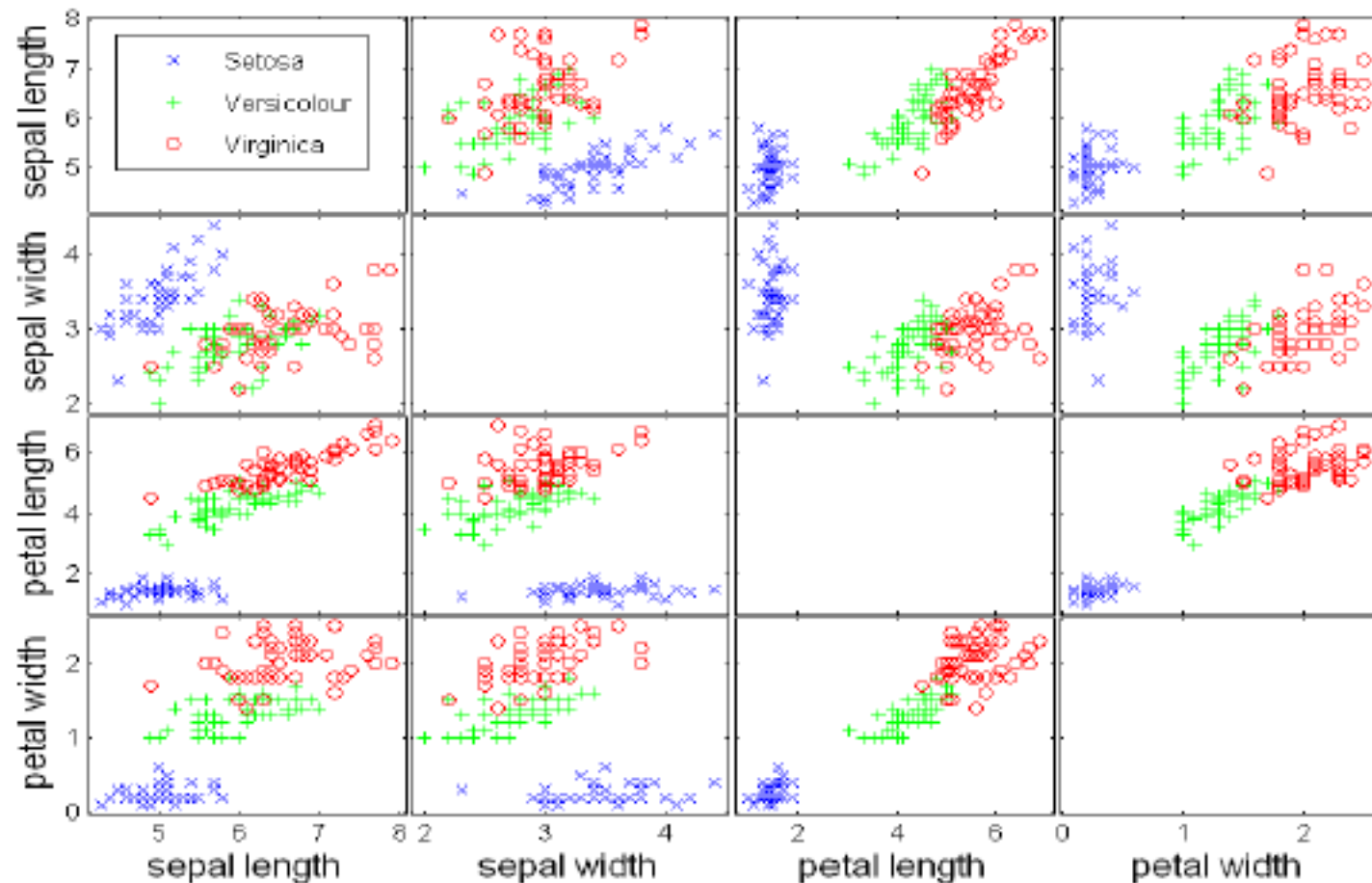
- Diagramas para visualizar dados multivariados
  - Em particular, relação entre diferentes atributos
  - Alguns tipos de gráficos:
    - *Scatter plot*
    - *Bags plot*
    - Faces de *Chernoff*
    - *Star plots*
    - *Heatmaps*

# Scatter plot

- Ilustra a correlação linear entre dois atributos
  - Cada objeto, considerando apenas dois de seus atributos, é associado a uma posição em um plano
    - Valores dos atributos definem a sua posição
    - Valores são inteiros ou reais
  - Matrizes de scatter plot: relacionamento de vários atributos

# Scatter plot

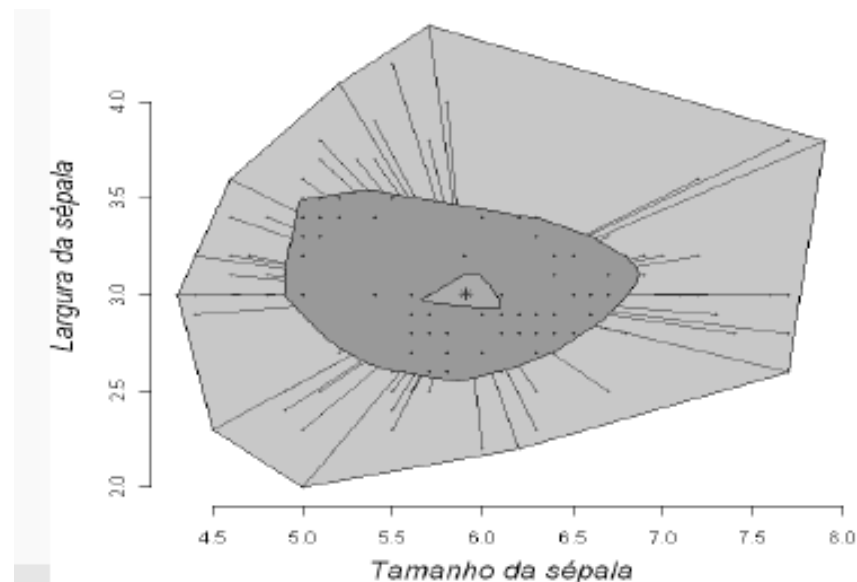
- Ex: conjunto de dados da iris



# Bagplot

## ■ Generalização bivariada do *boxplot*

- Apresenta, em uma mesma figura, o *boxplot* de dois atributos
  - Cada eixo pode ser considerado um *boxplot* de um dos atributos
- Ex: conjunto de dados iris

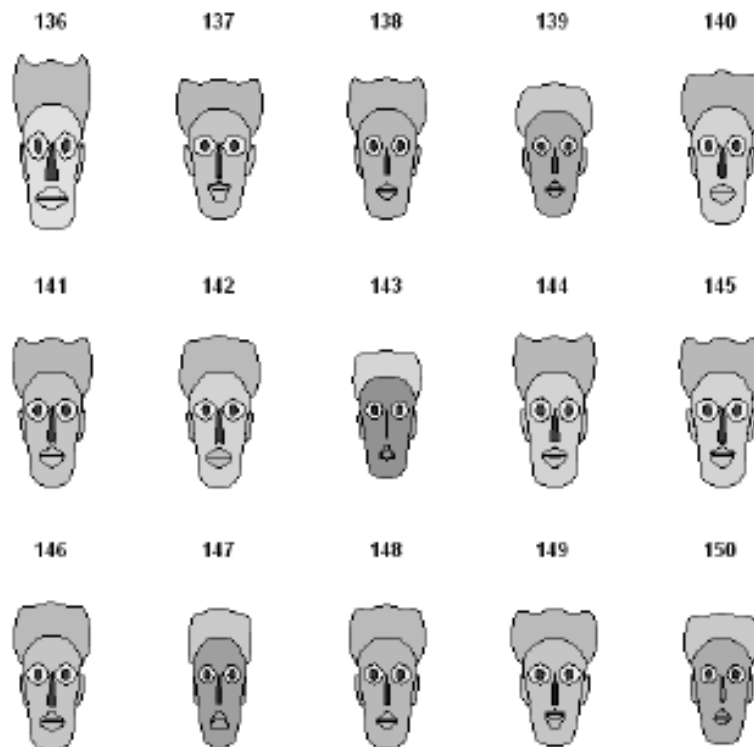


# Diagrama de Chernoff

- Mapeia valores dos atributos para imagens mais familiares: faces
  - Cada objeto (dado) é representado por uma face
  - Cada atributo é associado a uma ou mais características da face
    - Ex: altura e largura da cabeça, da boca, etc
- Baseia-se na habilidade humana de distinguir faces

# Diagrama de Chernoff

- Ex: base de dados iris



Tamanho da sépala  
representado por  
altura da face,  
largura da boca,  
altura do cabelo e  
largura do nariz

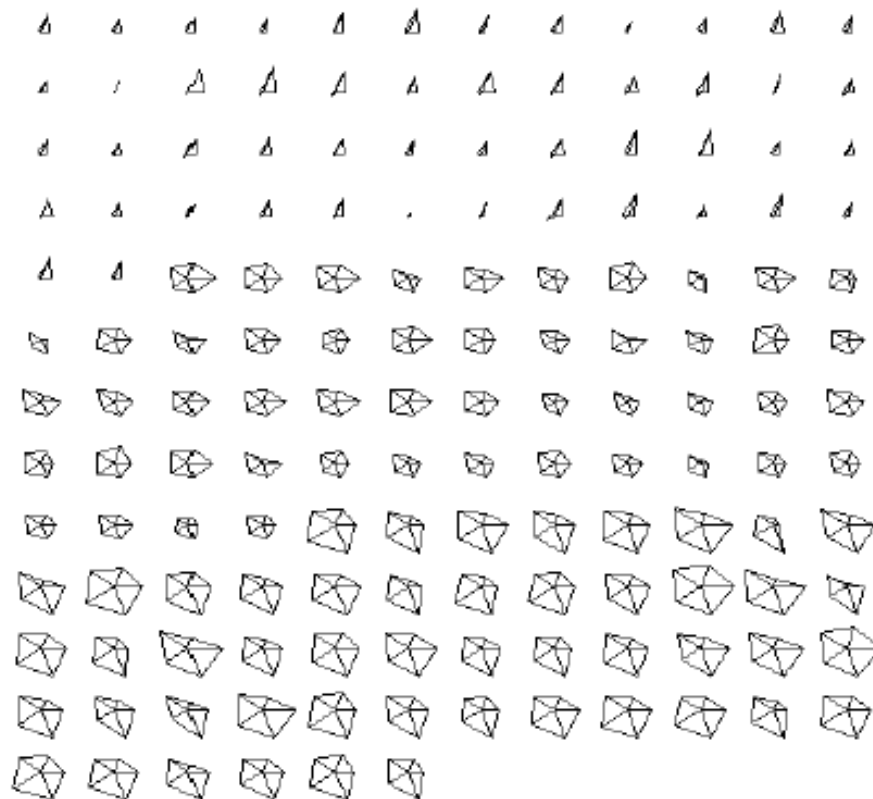
# Star plot

- Desenha uma figura geométrica para cada objeto
  - Normalmente um polígono
  - Cada linha do polígono corresponde a um dos atributos
    - Tamanho da linha é proporcional ao valor do atributo
    - Quanto mais atributos, mais o polígono se assemelha a estrela
    - Valores de atributos semelhantes deformam a estrela



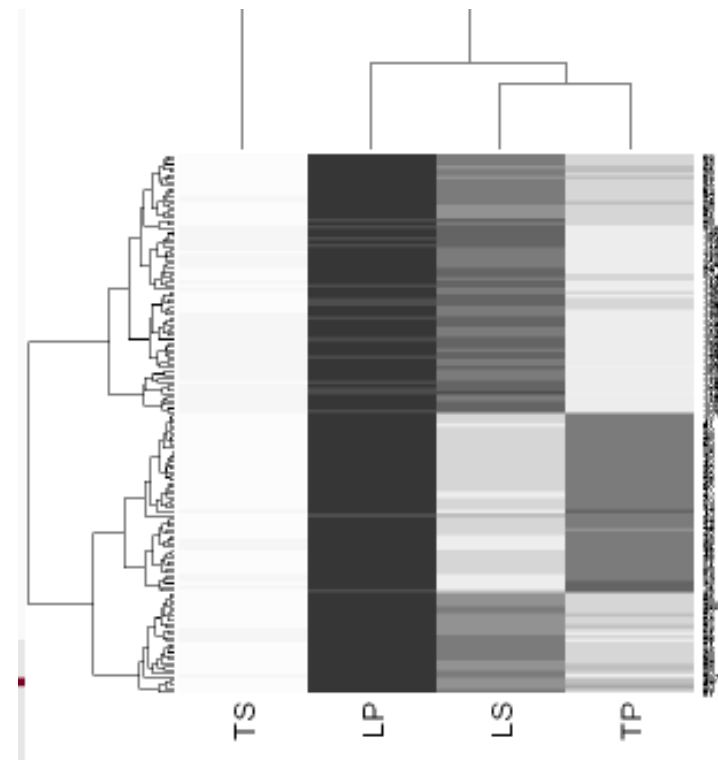
# Star plot

- Ex: conjunto de dados da iris



# Heatmap

- Representa a relação entre exemplos e as classes
  - Agrupamento hierárquico (dendograma)
    - Auxilia a verificar tendências nos dados
    - Ex: conjunto de dados iris





# Referências

- Material de aula da profa. Dra. Ana Carolina Lorena e o Livro Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina