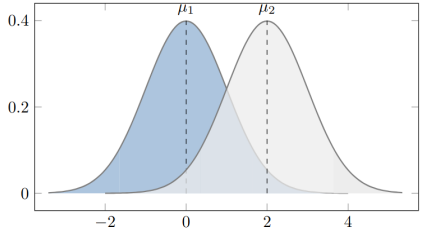
Análise exploratória de dados

Parte 7

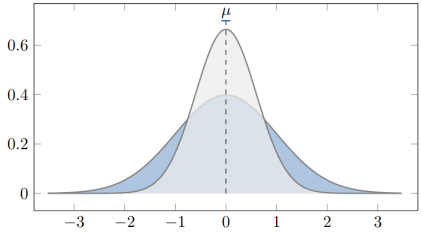
Prof.: Eduardo Vargas Ferreira

Medidas de posição e dispersão

Medidas de posição



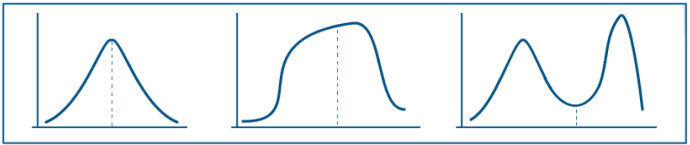
Medidas de dispersão



Medidas de Dispersão

Medidas de Dispersão

- O resumo de um conjunto de dados por medidas de posição esconde toda a informação sobre a variabilidade das observações.



- Um critério freqüentemente usado para esse fim é aquele que mede a dispersão dos dados em torno de sua média. O mais usado é a **variância amostral**:

$$Var(X) = \sum_{i=1}^{d} \frac{n_i \cdot (x_i - \bar{x})^2}{n}$$

Exemplo: número de filhos

- Voltando ao exemplo sobre a frequência dos funcionários segundo o nº de filhos, vimos que $\bar{x} = 1,65$.

Nº de filhos x_i	**Frequência** n_i	**Proporção** f_i
0	4	0.20
1	5	0.25
2	7	0.35
3	3	0.15
5	1	0.05
Total	20	1.00

$$Var(X) = \sum_{i=1}^{d} \frac{n_i \cdot (x_i - \bar{x})^2}{n} = \frac{4}{20} \cdot (0 - 1,65)^2 + \frac{5}{20} \cdot (1 - 1,65)^2 + \ldots + \frac{1}{20} \cdot (5 - 1,65)^2$$

$$= \sum_{i=1}^{d} f_i \cdot (x_i - \bar{x})^2 = 1,52$$

Exemplo: número de filhos

- Voltando ao exemplo sobre a frequência dos funcionários segundo o nº de filhos, vimos que $\bar{x} = 1,65$.

Nº de filhos x_i	**Frequência** n_i	**Proporção** f_i
0	4	0.20
1	5	0.25
2	7	0.35
3	3	0.15
5	1	0.05
Total	20	1.00

Variância amostral

$$Var(X) = \sum_{i=1}^{d} f_i \cdot (x_i - \bar{x})^2$$

$$= 1,52 \text{ filhos}^2$$

Desvio padrão

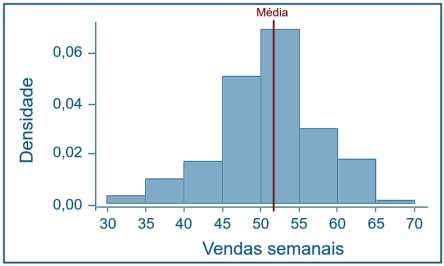
$$dp(X) = \sqrt{Var(X)}$$

$$= 1,23 \text{ filhos}$$

Exemplo: vendas semanais

- Os dados representam as vendas semanais de vendedores de gêneros alimentícios:

Vendas semanais	Nº de vendedores
$30 \vdash 35$	2
$35 \vdash 40$	10
$40 \vdash 45$	18
$45 \vdash 50$	50
$50 \vdash 55$	70
$55 \vdash 60$	30
$60 \vdash 65$	18
$65 \vdash 70$	2



1. Calcule o desvio padrão da amostra?

$$Var(X) = \frac{2 \cdot (32.5 - 51,2)^2 + 10 \cdot (37.5 - 51,2)^2 + \ldots + 2 \cdot (67.5 - 51,2)^2}{200}$$

$$= 43.81.$$

$$dp(X) = \sqrt{43.81} = 6.61.$$

Medidas complementares

Coeficiente de variação (CV)

- O CV é uma medida de **dispersão relativa** definida como a razão entre o desvio padrão e a média. Um valor superior a 50% sugere alta dispersão, o que indica heterogeneidade dos dados.

$$CV = \frac{dp}{\bar{x}}$$

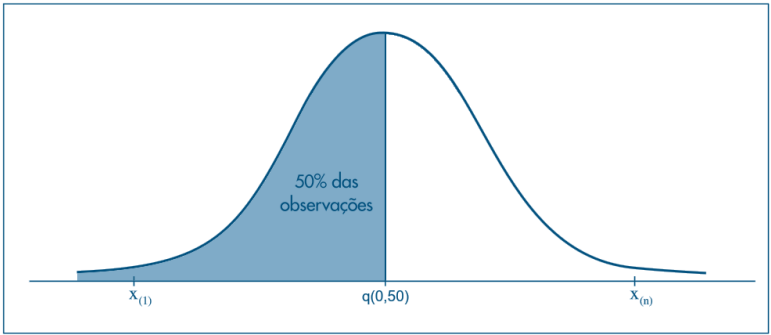
- Quanto mais próximo de zero, mais homogêneo são os dados e mais representativa será a média.

Grupo	Valores	Média	Desvio padrão
A	$10, 20, 30$	20	10
B	$10000, 10010, 10020$	10010	10

$$CV_A = \frac{10}{20} = 0,5 \qquad CV_B = \frac{10}{10010} \approx 0.0009$$

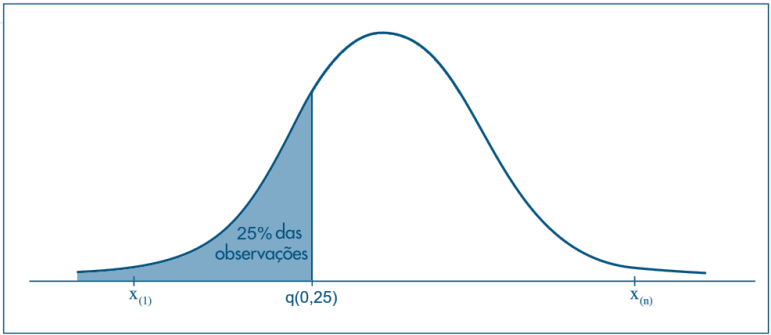
Quantis empíricos

- Vimos que a **mediana** deixa metade dos dados abaixo desse valor e a outra metade acima. Mas, podemos definir qualquer separação através dos **quantil de ordem p**, com $0 < p < 1$.



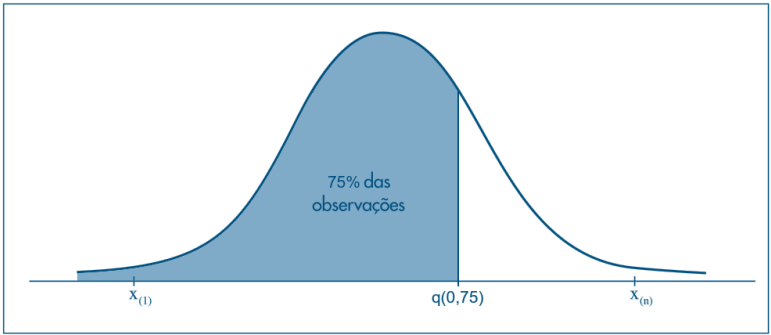
Quantis empíricos

- Vimos que a **mediana** deixa metade dos dados abaixo desse valor e a outra metade acima. Mas, podemos definir qualquer separação através dos **quantil de ordem p**, com $0 < p < 1$.



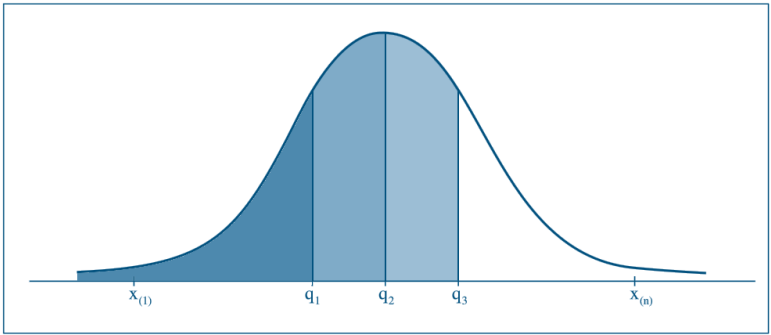
Quantis empíricos

- Vimos que a **mediana** deixa metade dos dados abaixo desse valor e a outra metade acima. Mas, podemos definir qualquer separação através dos **quantil de ordem p**, com $0 < p < 1$.



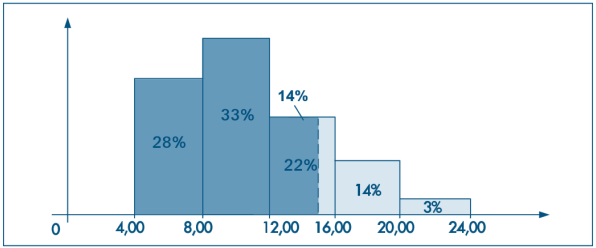
Quantis empíricos

- Vimos que a **mediana** deixa metade dos dados abaixo desse valor e a outra metade acima. Mas, podemos definir qualquer separação através dos **quantil de ordem p**, com $0 < p < 1$.



Exemplo: valores agrupados

- Se os dados estiverem agrupados em classes, podemos obter os quantis usando o histograma. P. ex.:



Mediana

$$\frac{q(0,50) - 8,00}{22\%} = \frac{12,00 - 8,00}{33\%}$$

$$q(0,50) = 8,00 + 2,67 = 10,67.$$

3º quartil

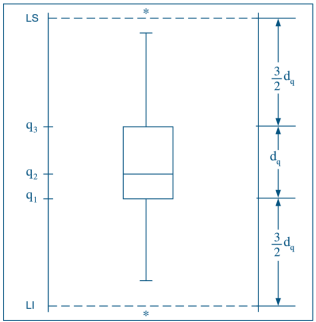
$$\frac{q(0,75) - 12,00}{14\%} = \frac{16,00 - 12,00}{22\%}$$

$$q(0,75) = 14,55.$$

Esquema dos cinco números

- Os valores $x_{(1)}, q_1, q_2, q_3$ e $x_{(n)}$ formam o **esquema dos cinco números**.

- As informações contida neste esquema pode ser traduzida através do *box-plot*.



- **Intervalo Interquartil:**

$$d_q = q_3 - q_1$$

- **Limite superior:**

$$LS = q_3 + (1,5) \cdot d_q$$

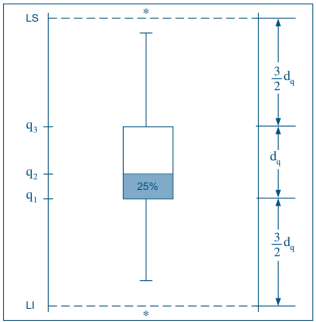
- **Limite inferior:**

$$LI = q_1 - (1,5) \cdot d_q$$

Esquema dos cinco números

- Os valores $x_{(1)}, q_1, q_2, q_3$ e $x_{(n)}$ formam o **esquema dos cinco números**.

- As informações contida neste esquema pode ser traduzida através do *box-plot*.



- **Intervalo Interquartil:**

$$d_q = q_3 - q_1$$

- **Limite superior:**

$$LS = q_3 + (1,5) \cdot d_q$$

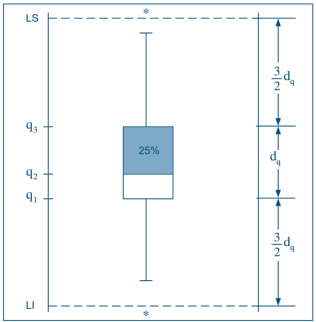
- **Limite inferior:**

$$LI = q_1 - (1,5) \cdot d_q$$

Esquema dos cinco números

- Os valores $x_{(1)}, q_1, q_2, q_3$ e $x_{(n)}$ formam o **esquema dos cinco números**.

- As informações contida neste esquema pode ser traduzida através do *box-plot*.



- **Intervalo Interquartil:**

$$d_q = q_3 - q_1$$

- **Limite superior:**

$$LS = q_3 + (1,5) \cdot d_q$$

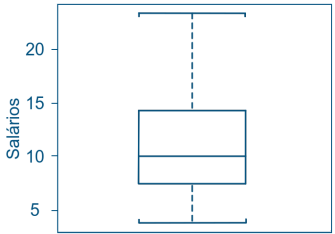
- **Limite inferior:**

$$LI = q_1 - (1,5) \cdot d_q$$

Exemplo: salário dos empregados

- Os dados abaixo referem-se ao salário de 36 funcionário de uma empresa.

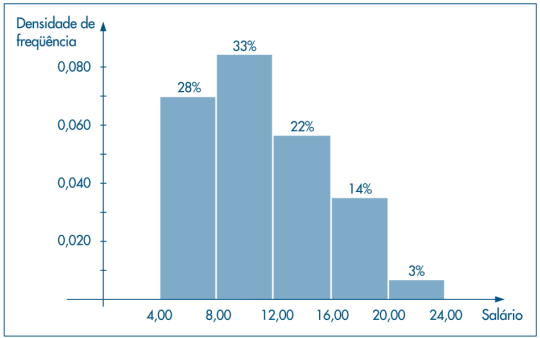
Classe de salários	Frequência
$4,00 \vdash 8,00$	10
$8,00 \vdash 12,00$	12
$12,00 \vdash 16,00$	8
$16,00 \vdash 20,00$	5
$20,00 \vdash 24,00$	1
Total	36



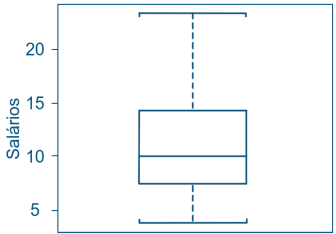
- Neste caso, observa-se uma distribuição assimétrica à direita.

Exemplo: salário dos empregados

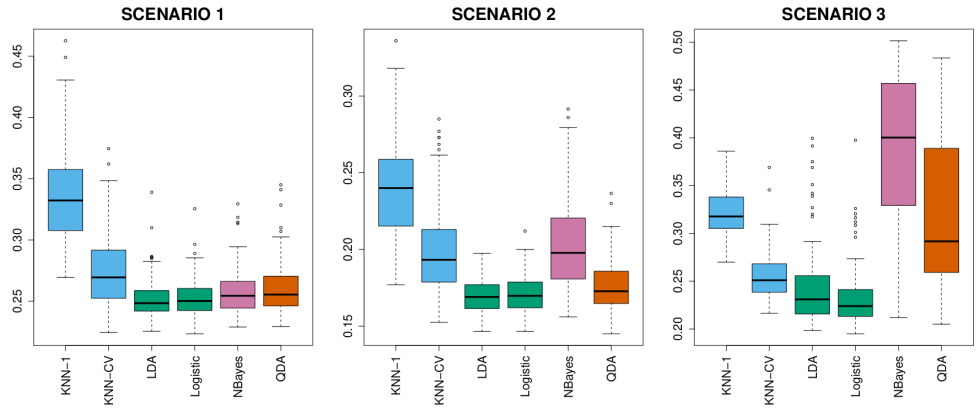
- Os dados abaixo referem-se ao salário de 36 funcionário de uma empresa.



Fonte: Estatística Básica (Bussab e Morettin, 2017)



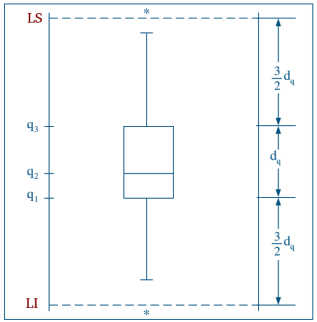
Exemplo: comparação de algoritmos de Machine Learning

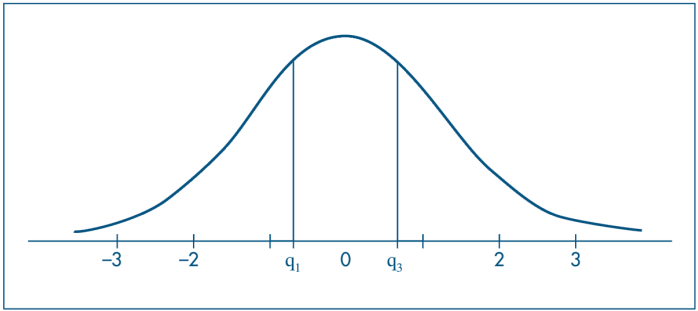


Fonte: An Introduction to Statistical Learning.

Outliers

- Para entendermos os motivos de se utilizar os limites **LI** e **LS** para definir as observações atípicas (*outliers*), considere a distribuição abaixo:

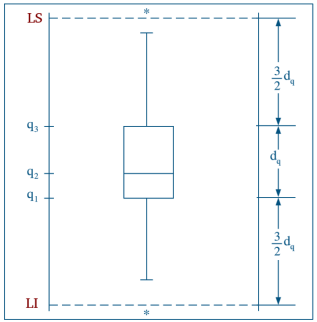


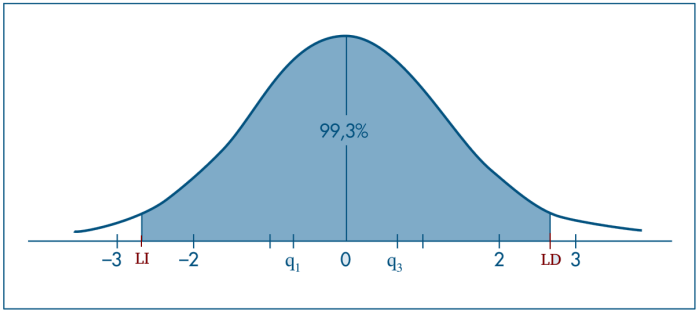


- A área entre **LI** e **LS** representa 99,3% da distribuição.

Outliers

- Para entendermos os motivos de se utilizar os limites **LI** e **LS** para definir as observações atípicas (*outliers*), considere a distribuição abaixo:





- A área entre **LI** e **LS** representa 99,3% da distribuição.

Referências

- Bussab, WO; Morettin, PA. Estatística Básica. São Paulo: Editora Saraiva, 2006 (5ª Edição).
- Magalhães, MN; Lima, ACP. Noções de Probabilidade e Estatística. São Paulo: EDUSP, 2008.