Insper

Ciência dos Dados

Aula 04

Análise Exploratória dos Dados

Objetivos de Aprendizagem

Os alunos devem ser capazes de:

- Desenvolver medidas que gerem informações para interpretação de variáveis quantitativas.
- Interpretar o comportamento de uma variável quantitativa a partir dos formatos de um histograma e/ou um box-plot.
- Comparar cenários, a partir dos gráficos e medidas calculadas, para tomada de decisão.

Medidas de posição

Notação

Amostra de n observações da variável X:

$$x_1, x_2, \ldots, x_n$$

Amostra **ordenada** de n observações da variável X:

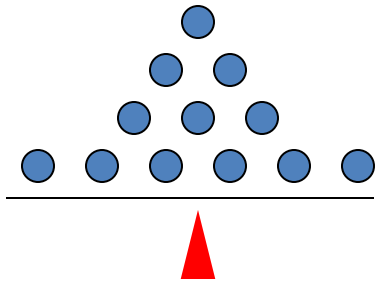
$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Mínimo = $x_{(1)}$

Máximo = $x_{(n)}$

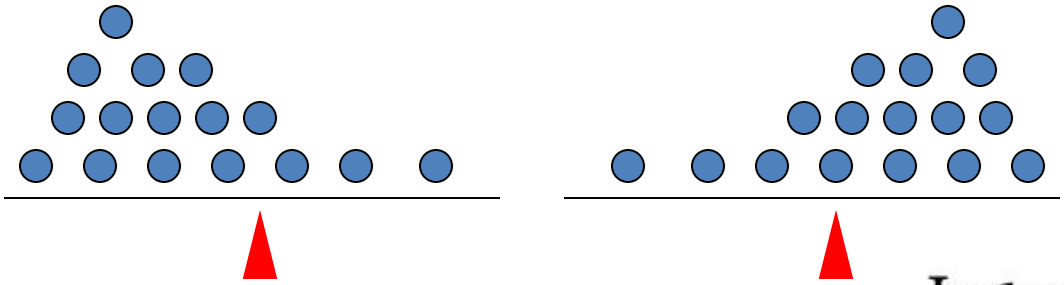
Média Aritmética

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$



Valores aberrantes

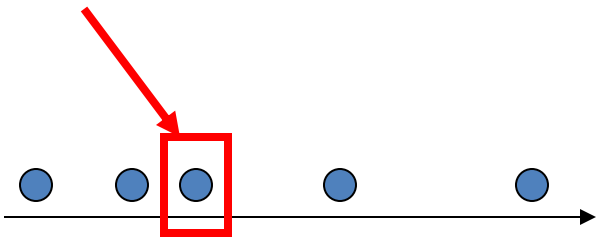
Assimetrias

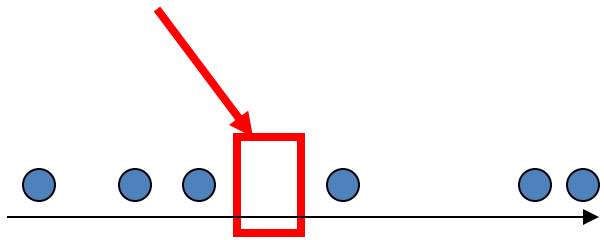


Mediana

Mediana é o valor que divide um conjunto de dados ordenados ao meio. Em outras palavras, existe uma quantidade igual de valores menores e maiores que um dado valor

$$\text{md}(X) = \begin{cases} \text{valor da } \left(\frac{n+1}{2}\right)^{\underline{a}} \text{ observação ordenada; se n é ímpar} \\ \text{valor médio entre a } \left(\frac{n}{2}\right)^{\underline{a}} \text{ e a } \left(\frac{n}{2}+1\right)^{\underline{a}} \text{ observações ordenadas; se n é par} \end{cases}$$





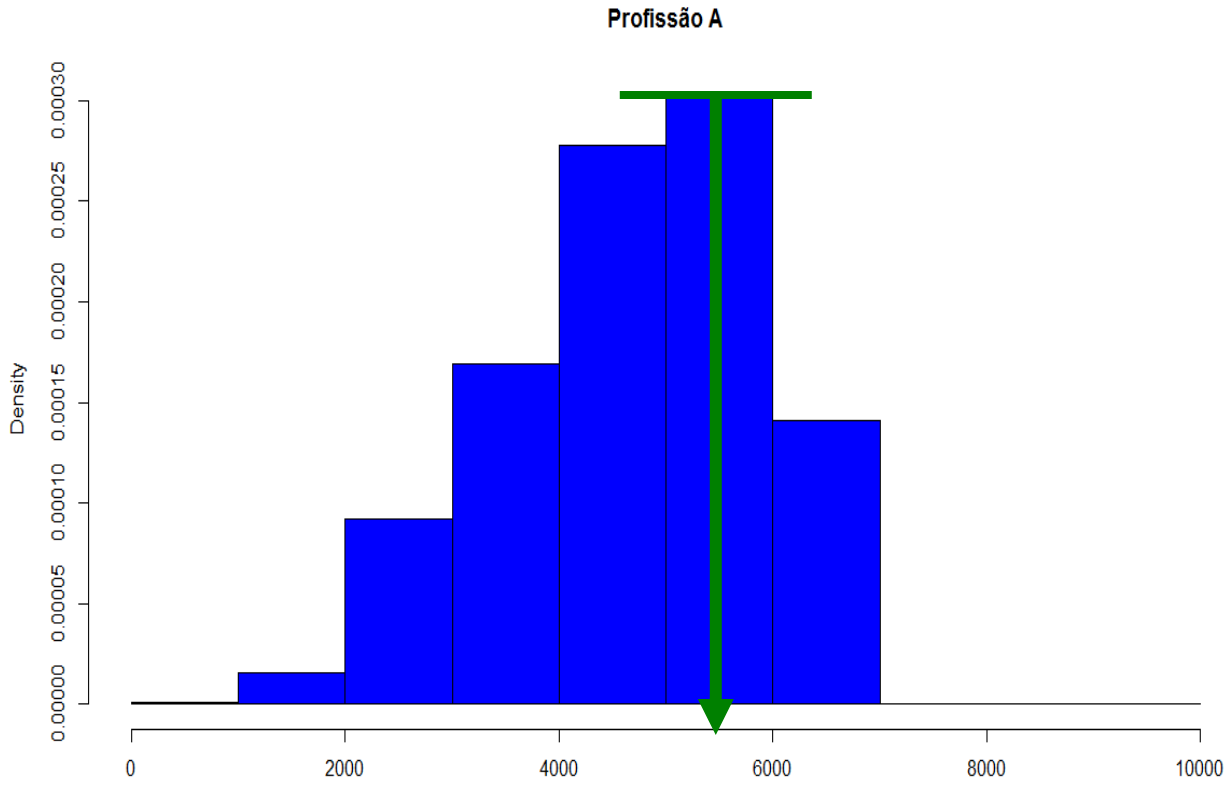
Moda

Moda de um conjunto de dados, representado por **mo(x)**, é o valor que ocorre com maior frequência.

- Quando dois valores ocorrem com a mesma maior frequência, cada um é uma moda e o conjunto de dados é chamado de **bimodal**.
- Quando mais de dois valores ocorrem com a mesma maior frequência, cada um é uma moda é o conjunto de dados é **multimodal**.
- Quando nenhum valor se repete, dizemos que **não há moda (amodal)**.

Moda
(escolhendo a classe modal)

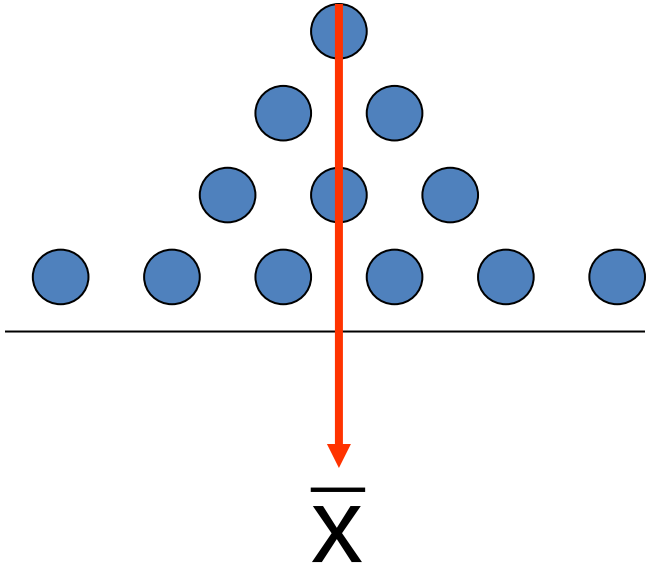
mo(X) = classe com maior densidade



mo(X)

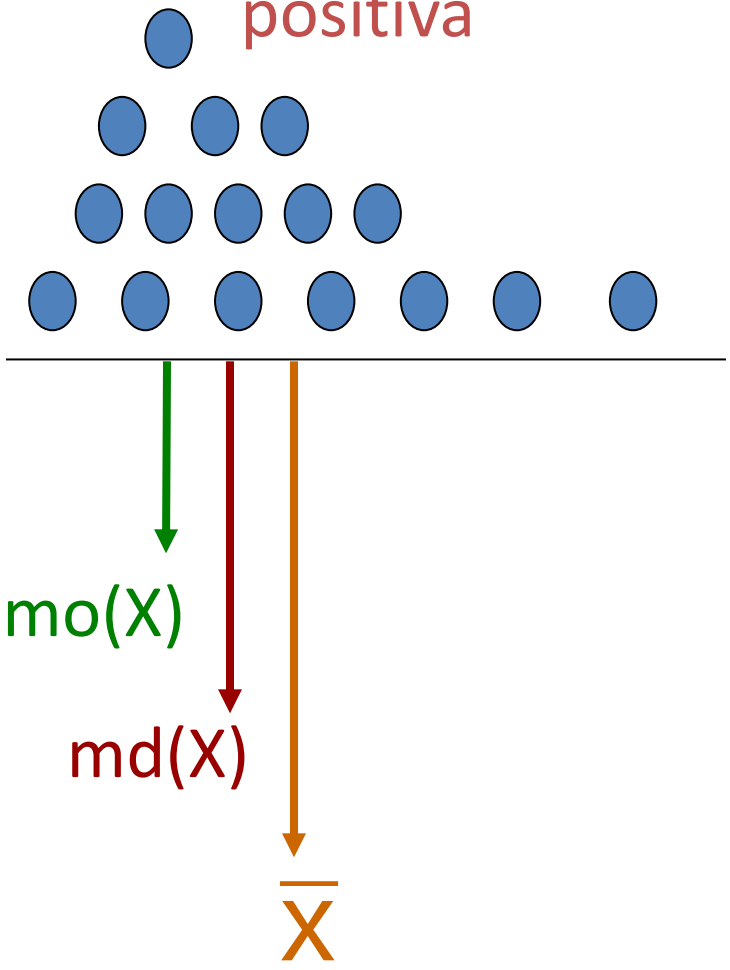
Formato da distribuição e medidas de tendências

Distribuições Simétricas

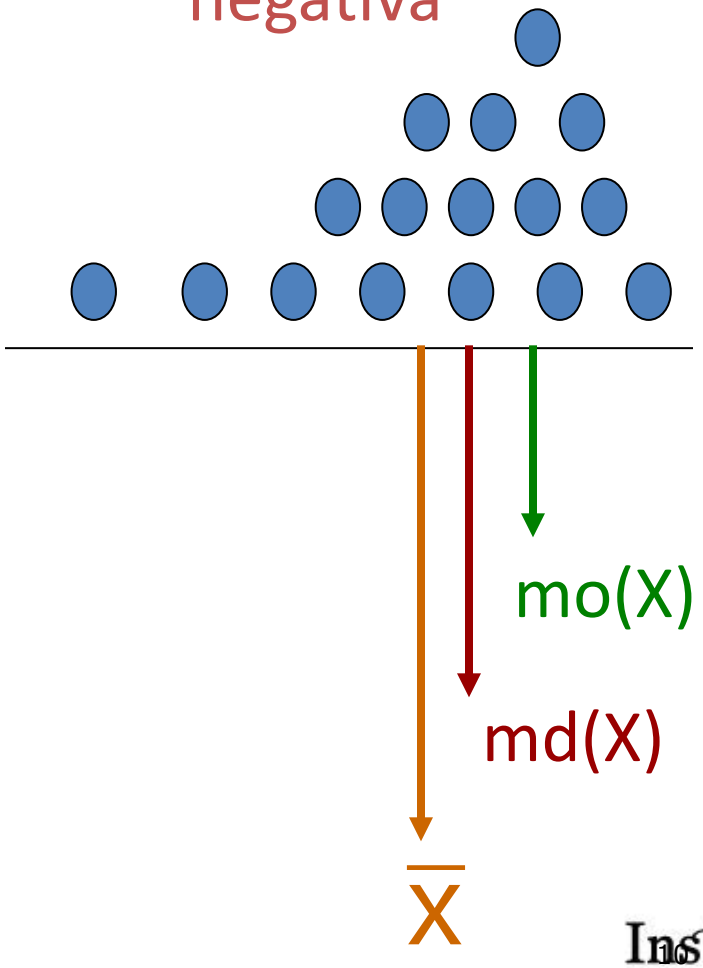


Formato da distribuição e medidas de tendências

Assimetria à direita ou positiva



Assimetria à esquerda ou negativa



Insper

Medidas de dispersão

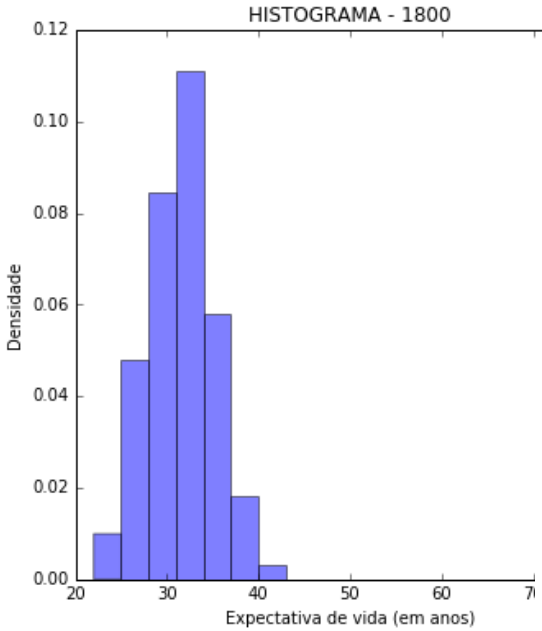
Expectativa de vida

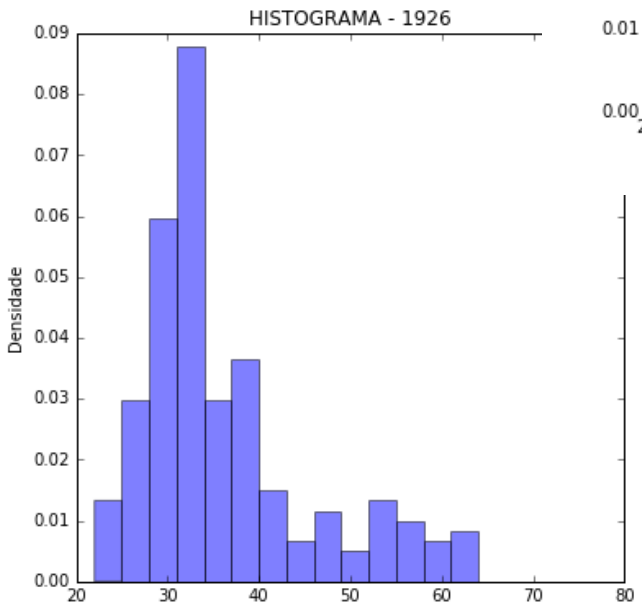
Expectativa de vida (em anos) dos anos 1800, 1926 e 2000.

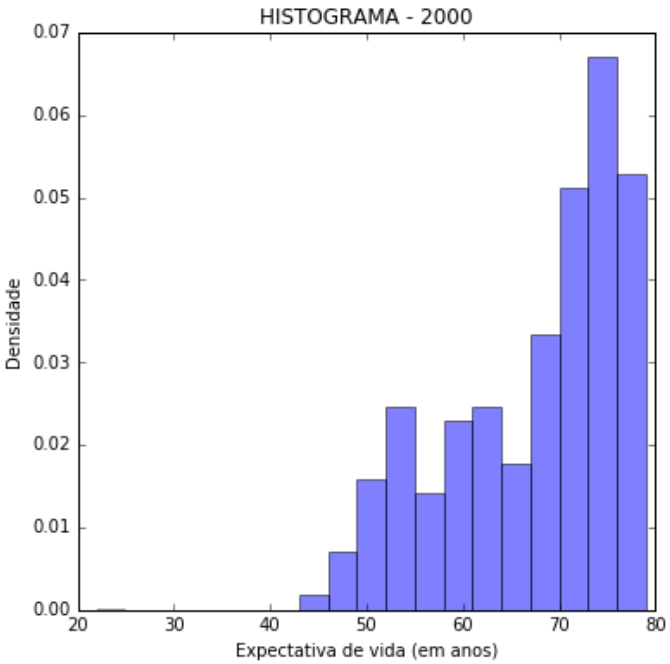
Medidas Resumo	1800	1926	2000
Média	31,5	36,3	68,0
Mediana	31,8	32,8	71,2
Desvio Padrão	3,76	9,59	9,21
Variância	14,16	91,99	84,86
Desvio Médio Absoluto	2,98	7,40	7,79
Tamanho amostral	201	201	201

Expectativa de vida

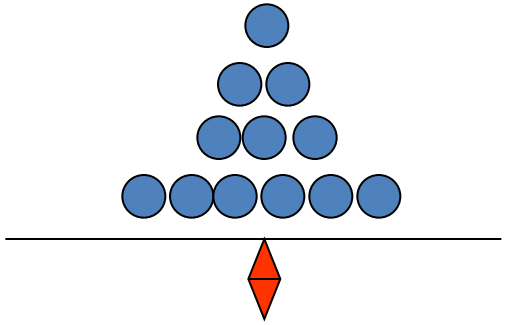
Expectativa de vida (em anos) dos anos 1800, 1926 e 2000.





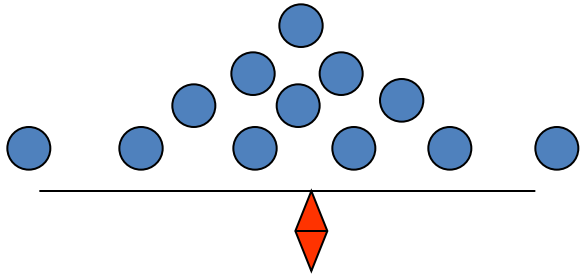


Medidas de Dispersão baseadas em distâncias a uma medida de tendência central



Baixa variabilidade

As observações estão próximas à medida de tendência central



Alta variabilidade

As observações estão mais distantes da medida de tendência central

Medidas de dispersão

Variância: a variância da amostra é a média das diferenças ao quadrado entre cada uma das observações e a média do conjunto.

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

Características:

- Um dos problemas de usar a variância como medida de dispersão é o fato de sua unidade não ser a mesma unidade em que a variável foi medida (os valores dos dados estão elevados ao quadrado). A solução é extrair a raiz quadrada positiva da variância, já que, com isso, se volta à unidade original da variável.

Medidas de dispersão

Desvio Padrão: o desvio padrão de um conjunto de valores amostrais é uma medida da variação dos valores em torno da média. É uma espécie de desvio médio dos valores em relação à média aritmética.

$$dp(X) = \sqrt{var(X)} = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ n}$$

Características:

- É positivo. É zero apenas quando todos os valores dos dados são o mesmo número.
- O valor do desvio padrão pode crescer dramaticamente com a inclusão de um ou mais *outliers.*
- A unidade de medida é a mesma da variável X.

Desvio Médio Absoluto

Desvio Médio Absoluto: é a distância média dos dados até a média (aritmética).

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

Características:

Na prática, não é uma medida muito utilizada para resumir a variabilidade de um conjunto de dados. Isso ocorre porque a função módulo não é diferenciável o que cria dificuldades para alguns métodos de inferência estatística.

Insper

Insper

Percentis e Boxplot

Magalhães e Pedroso de Lima, Cap 1

Percentil ou Quantil

Amostra ordenada

p% menores observações

(100-p)% maiores observações

Quantil ou Percentil de ordem p (0<p<100): é o valor que divide o conjunto de dados ordenado em 2 partes: uma delas com p% dos menores valores e a outra com (100-p)% dos maiores valores.

Insper

Expectativa de vida – Percentis

Expectativa de vida (em anos) dos anos 1800, 1926 e 2000.

Ordem	1800	1926	2000
Mínimo	23	23	46
10%	26	27	53
20%	29	30	60
30%	30	31	64
40%	31	32	69
50%	32	33	71
60%	32	35	73
70%	33	38	74
80%	35	41	76
90%	36	53	78
Máximo	43	63	83

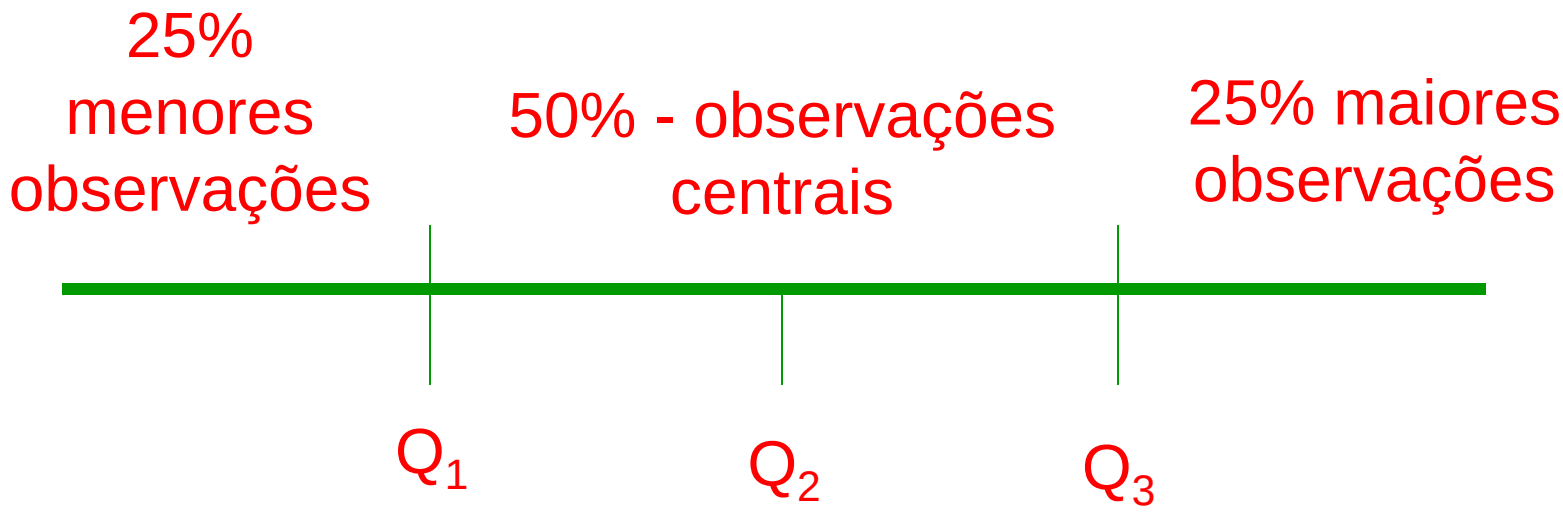
Life expectancy	1800	1926	2000
Brazil	32	31,99072	71,9

Comando Python:

np.percentile (variável, ordem entre 0 e 100) ou variável.quantile(ordem em 0 e 1)

Quartis

Amostra ordenada



Quartil: são valores que dividem o conjunto de dados ordenados em quatro partes iguais.

Cada parte contendo 25% dos dados.

Intervalo Interquartil: IQ = $Q_3 - Q_1$

Exercício: Dirigindo Bêbado

Excedendo o limite de 0,1 grama de álcool por litro de sangue:

paga multa, perde a carteira e tem carro apreendido.

Excedendo o limite de 0,6 grama de álcool por litro de sangue:

pode ser preso.

Os 14 motoristas abaixo foram parados em uma blitz policial.

0,32	**0,39**	**0,02**	**0,18**	**0,13**	**0,49**	**0,63**
0,08	**0,08**	**0,16**	**0,08**	**0,26**	**0,16**	**0,25**

Dado que a atual lei proíbe dirigir com níveis acima de 0,1, parece que esses níveis estão acima do permitido?

Encontre os quartis, calcule o IQ e discuta a assimetria.

Exercício: Dirigindo Bêbado

Conjunto de Dados:	0,32	0,39	0,02	0,18	0,13	0,49	0,73
	0,08	0,08	0,16	0,08	0,26	0,16	0,25

Dados Ordenados:	0,02	0,08	0,08	0,08	0,13	0,16	0,16
	0,18	0,25	0,26	0,32	0,39	0,49	0,73

Q1 = 0,08

Q2 = 0,17

Q3 = 0,32

IQ = Q3 – Q1 = 0,32 – 0,08 = 0,24

Boxplot

O **boxplot** é uma figura que possibilita visualizar várias características de um conjunto de dados como:

- as de tendência central (mediana)
- de posição (primeiro quartil e terceiro quartil)
- de dispersão (intervalo entre quartis)
- de assimetria
- pode **identificar os valores considerados como possíveis extremos**.

Identificação de possíveis valores aberrantes

Um ponto (w) será considerado suspeito de ser aberrante se:

$w > LS = Q_3 + 1{,}5\ IQ$ (Limite Superior), ou

$w < LI = Q_1 - 1{,}5\ IQ$ (Limite Inferior),

sendo $IQ = Q_3 - Q_1$ (intervalo interquartílico)

Propriedade: num modelo Normal, apenas 0,7% dos dados estão fora desses limites.

Limitação: em distribuições assimétricas a regra tende a identificar um número excessivo de valores suspeitos.

Boxplot

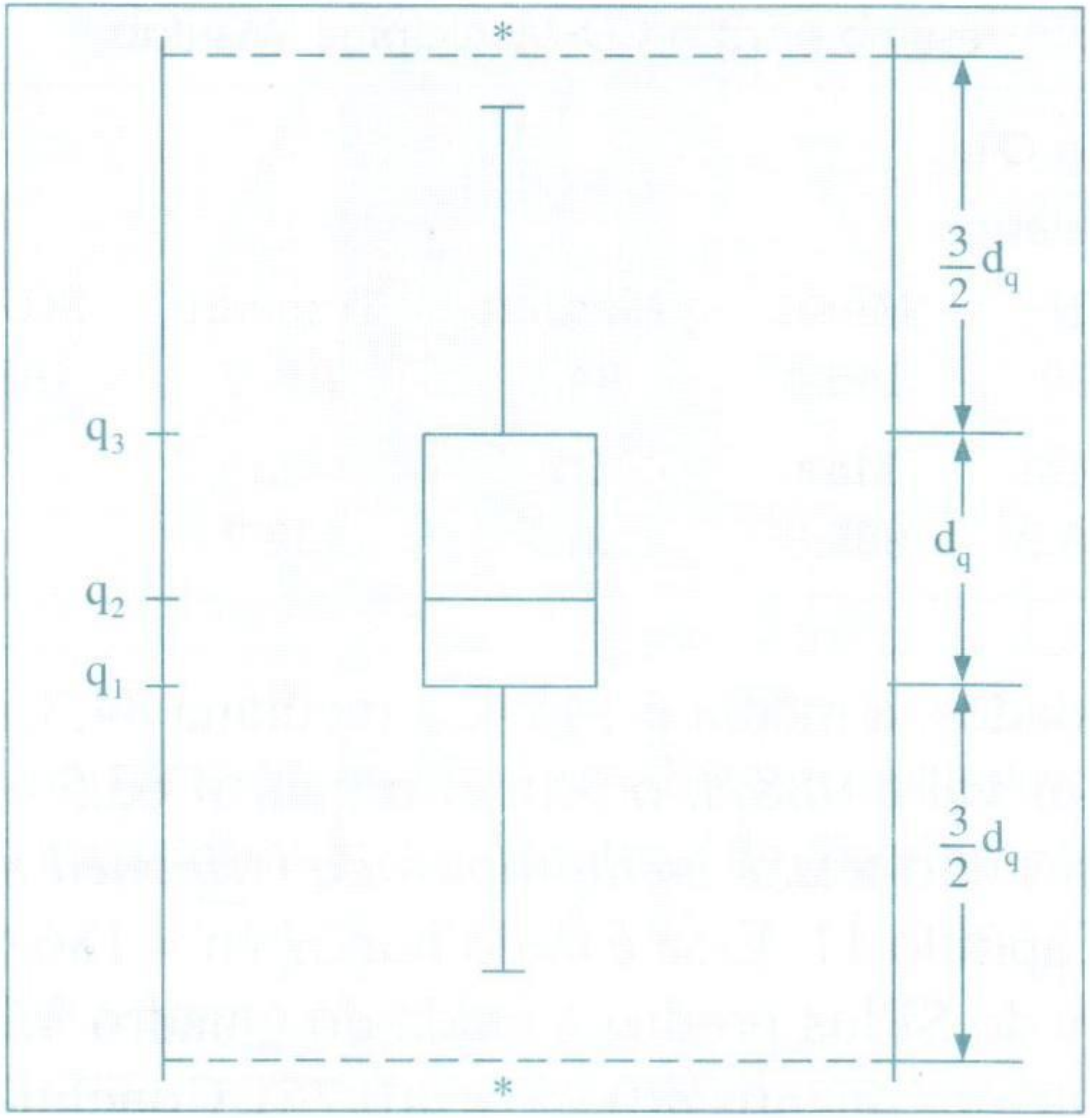
Caixa que contém 50% das observações centradas. Parte superior é Q3 e o inferior é Q1. Mediana Q2 está dentro da caixa. (IQ = Q3 - Q1)

Linha acima da caixa segue até o ponto mais remoto que não exceda o limite superior (LS = Q3 + 1,5 IQ)

Linha abaixo da caixa segue até o ponto mais remoto mas maior do que o limite inferior (LI = Q1 - 1,5 IQ)

Possível ponto extremo ou aberrante (“outlier”) – aqueles acima do LS ou abaixo do LI

Boxplot



$d_q = IQ$

Fonte: Bussab e Morettin

Exercício: Dirigindo Bêbado

Conjunto de Dados:							
	0,32	0,39	0,02	0,18	0,13	0,49	0,73
	0,08	0,08	0,16	0,08	0,26	0,16	0,25

Dados Ordenados:							
	0,02	0,08	0,08	0,08	0,13	0,16	0,16
	0,18	0,25	0,26	0,32	0,39	0,49	0,73

Q1 = 0,08

Q2 = 0,17

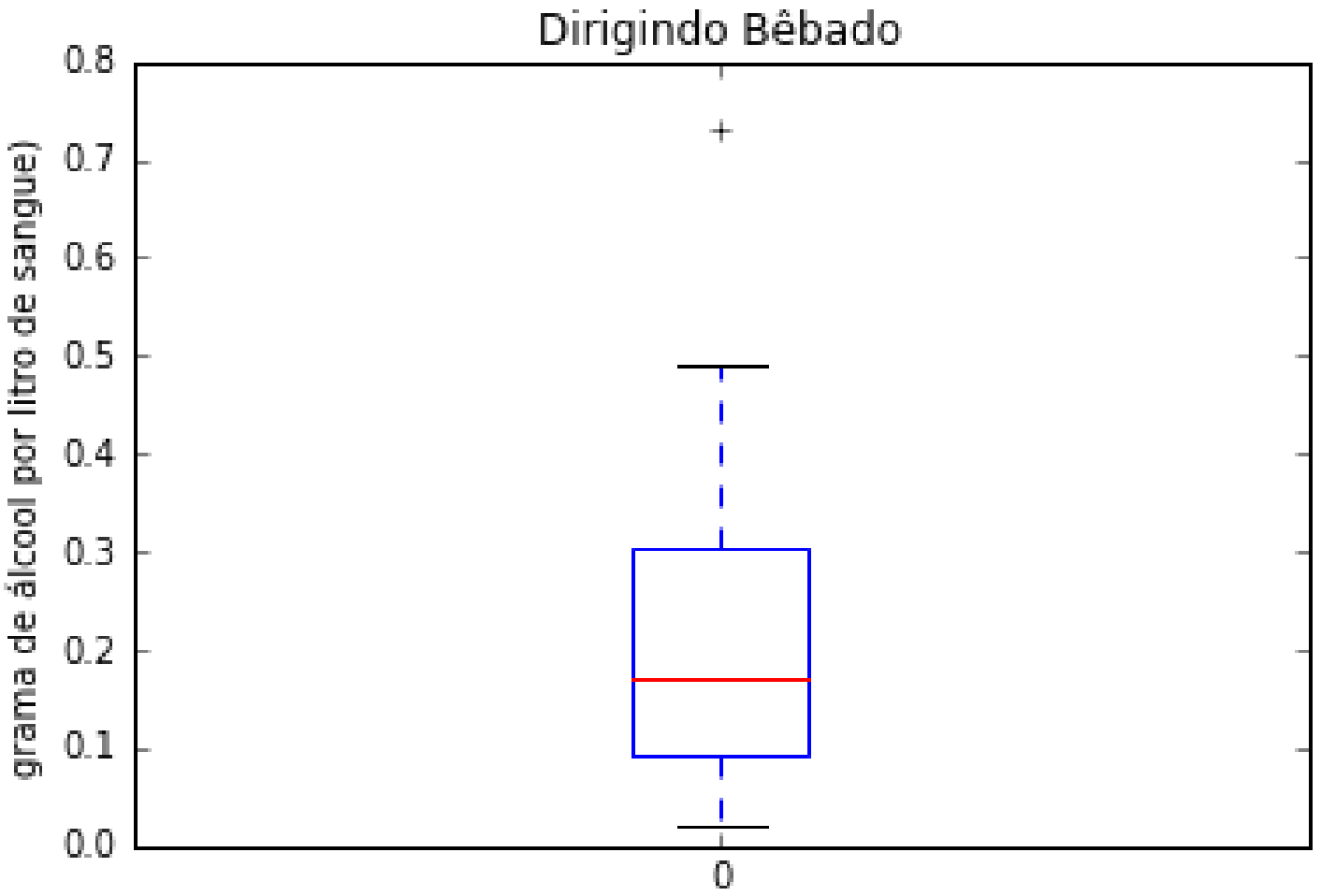
Q3 = 0,32

LI = 0,08 – 1,5 (0,32 – 0,08) = -0,28

LS = 0,32 + 1,5 (0,32 – 0,08) = 0,68

IQ = Q3 – Q1 = 0,32 – 0,08 = 0,24

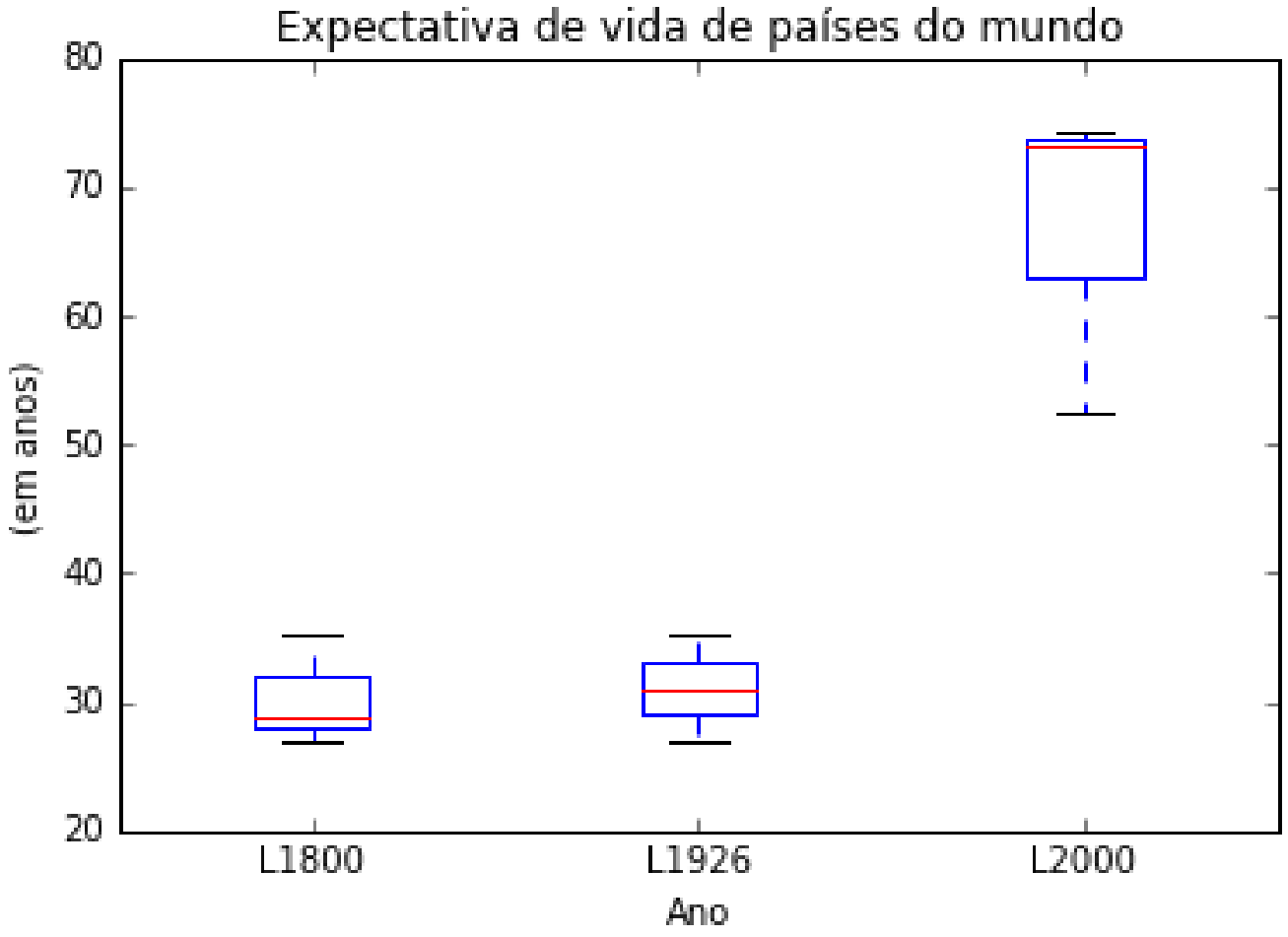
Boxplot



Comando Python:

variaveis.plot (kind='box') ou variáveis.plot.box( )

Boxplot



Comando Python:

variaveis.plot (kind='box') ou variáveis.plot.box( )

Atividade com Salários... (variável quantitativa)

?? minutos:

Análise descritiva de salários de quatro profissões diferentes.

Arquivo:

`Aula04 Atividade Variáveis Quantitativas com Salarios.ipynb`

Preparo para próxima aula

Os alunos devem se preparar com:

1. Leitura prévia necessária: Montgomery et al (5ª. Edição) - Seção 2.6: Dados Multivariados.
2. **Fazer PréAula05.** Não é para entregar.
3. Acrescentar **pelo menos DUAS variáveis quantitativas** para melhor detalhar na recomendação do Sr. Gold ➔ **Check na terça**!

Fiquem atentos aos Avisos no Blackboard!