Métodos Estatísticos em Pesquisa Científica

Aula 02 - Parte 1/2

Paulo Justiniano Ribeiro Jr

Departamento de Estatística
Setor de Ciências Exatas
Transversais - PRPPG
Universidade Federal do Paraná

27 de março, 2024

Revisando os Temas

- Dados I : obtenção, amostragem, pesquisa, volume, ...
- Dados II: descrição, resumos, gráficos, análises, ...
- Probabilidades.
- Inferência: incerteza, população, amostra, testes, intervalos, ...
- Modelagem e métodos.

Textos de apoio

Capítulos 1 **a** 4

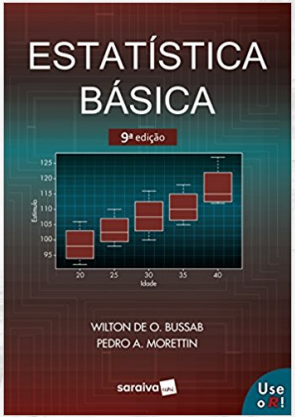


Figura 1. Bussab & Morettin

Cap. 1 **e** Cap. 4

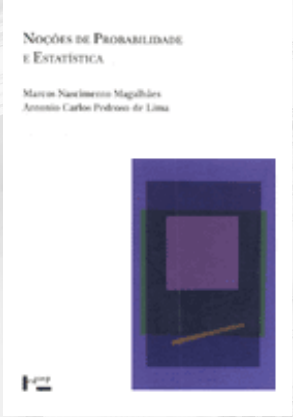
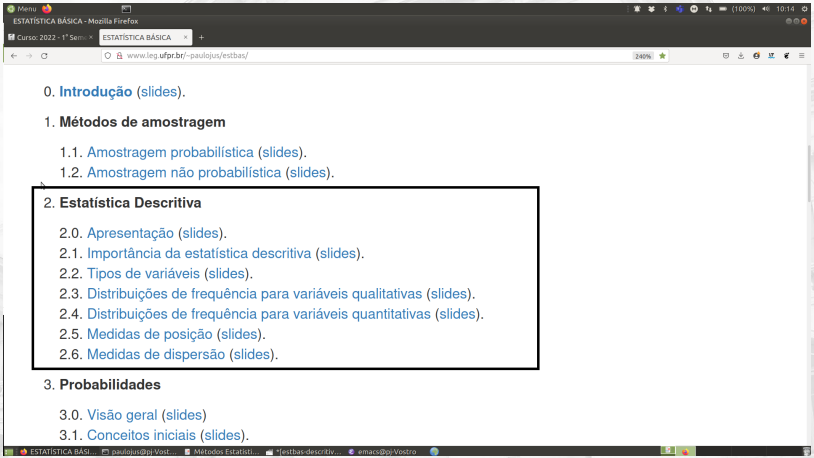


Figura 2. Magalhães & Lima

Tópicos equivalentes em uma **ENORME** diversidade de materiais.

Material de apoio

Curso de estatística básica (http://www.leg.ufpr.br/estbas)



0. **Introdução** (slides).

1. **Métodos de amostragem**

   1.1. Amostragem probabilística (slides).
   1.2. Amostragem não probabilística (slides).

2. **Estatística Descritiva**

   2.0. Apresentação (slides).
   2.1. Importância da estatística descritiva (slides).
   2.2. Tipos de variáveis (slides).
   2.3. Distribuições de frequência para variáveis qualitativas (slides).
   2.4. Distribuições de frequência para variáveis quantitativas (slides).
   2.5. Medidas de posição (slides).
   2.6. Medidas de dispersão (slides).

3. **Probabilidades**

   3.0. Visão geral (slides)
   3.1. Conceitos iniciais (slides).

Estatística descritiva

Análise exploratória de dados

- Resumir dados
- Apresentar de forma concisa e informativa
- Embasar discussões e argumentações
- Revelar padrões
- Descobrir possíveis anomalias
- Guiar *modelagem*
- ...

Análise descritiva

- Análises univariadas: descrição de perfil
- Análises bivariadas: explorando relações
- Análises multivariadas: relações complexas, gerais e confundimentos

Estruturas de dados

- Referência: tabela da dados
  - linhas: indivíduos
  - colunas: variáveis ou atributos
- Estruturas mais gerais (listas, bancos de dados, etc)
- Dados não estruturados

Resumindo dados

- Tabelas
- Gráficos
- Medidas estatísticas

Mas ... que tipos de tabelas/gráficos e medida podemos(ou devemos) usar?

Tipos de variáveis

Referência inicial:

- Qualitativas
  - Nominais
  - Ordinais
- Quantitativas
  - Discretas
  - Contínuas

Parte de assunto mais amplo de **visualização de dados** (tema de aula(s))

Tipos de variáveis

Variável
→ Característica.
→ Atributo.

Quantitativa
→ Expressa em números.
→ Variável **métrica**.

Discreta
→ **Finitos** valores em um intervalo.
→ # produtos vendidos/dia.
→ # gols/partida.
→ # passageiros/vôo.
→ # ovos quebrados/dúzia.

Contínua
→ **Infinitos** valores em um intervalo.
→ rendimento/mês de investimento.
→ consumo/mês de energia.
→ teor de argila do solo.

Qualitativa
→ Expressa em rótulos.
→ Variável **categórica**.

Nominal
→ Sem ordenação.
→ **Rótulo**.
→ preferência musical.
→ estado de origem.
→ signo do zoodíaco.

Ordinal
→ Com ordenação.
→ **Grau** ou **nível**.
→ grau de CNH.
→ grau de judô.
→ premiação olímpica.
→ nível de proficiência.

Exemplo 01: Dados *Milsa*

Dados do livro de Bussab & Morettin

Utilizados como exemplo em:
Material online pós-aula 01
(**Estatística Descritiva**)

Exemplo 02: Questionário

Tipos de variáveis com diversos resumos, tabelas e gráficos no:
Questionário do curso

Tipo	Exemplo	Posição	Dispersão
QL nominal	área de conhecimento	moda	?
QL ordinal	importância est.	moda mediana quantis	?
QT discreta QT contínua	número de artigos IMC	moda mediana quantis média(s)	amplitude, variância, desvio padrão, CV, amplitude interquartílica desvio médio

Exemplo 02: nominal

Grande área:
Biológicas: 53% , Exatas: 30% , Humanas 17%

Resumo:
moda = biológicas (mais frequente)
alguma medida de variabilidade?

Se tivéssemos
Biológicas: 83% , Exatas: 15% , Humanas 12%

A moda (posição) seria a mesma,
mas e a variabilidade (dispersão)?

Uma proposta: $V = -\sum_i p_i \log p_i$

Exemplo 02: variabilidade de nominal

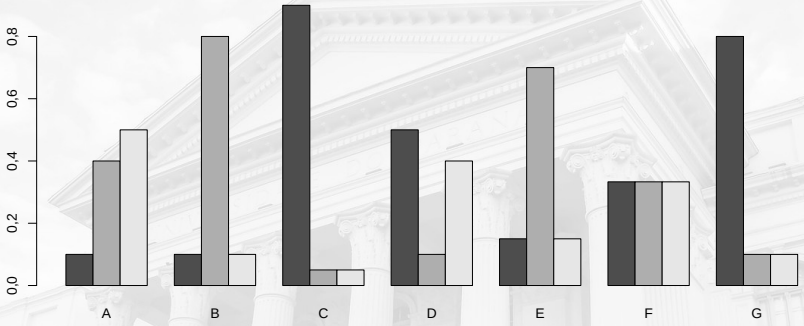


Figura 3. Diferentes possíveis proporções de três categorias

Variabilidades ($V = -\sum_i p_i \log p_i$):

A	B	C	D	E	F	G
0,943	0,639	0,394	0,943	0,819	1,099	0,639

Exemplo 03: Número de artigos

$$2, 0, 0, 3, 0, 1, 1, 0, 0, 0, 1, 4, 2, 0, 1$$

Os indivíduos da amostra publicaram uma média de 1.0 artigos, variando de 0 a 4.
Podemos ver o quanto cada um se afasta da média calculando os desvios:

$$1, -1, -1, 2, -1, 0, 0, -1, -1, -1, 0, 3, 1, -1, 0$$

Podemos avaliar se desviam muito ou pouco da média (*variabilidade/dispersão*) ...
...e representar isto em um número (*medida de dispersão*).
A soma dos desvios é sempre zero. (+/- se cancelam)
Podemos então pegar, por exemplo, os quadrados dos desvios

$$1, (-1)^2, (-1)^2, 2^2, (-1)^2, 0^2, 0^2, (-1)^2, (-1)^2, (-1)^2, 0^2, 3^2, 1^2, (-1)^2, 0^2$$

E fazer uma *média*[1] destes como uma outra *medida de dispersão*.
Esta é a variância = 1.6 e a sua raiz quadrada é o desvio padrão = 1.3.

O desvio padrão é mais "conveniente" pois possui a mesma unidade dos dados.

Notação

variável: número de artigos

vetor de dados: y

um dado individual: $y_i \;\; (y_1 = 2, y_2 = 0, y_3 = 0, \ldots, y_{14} = 0, y_{15} = 1)$

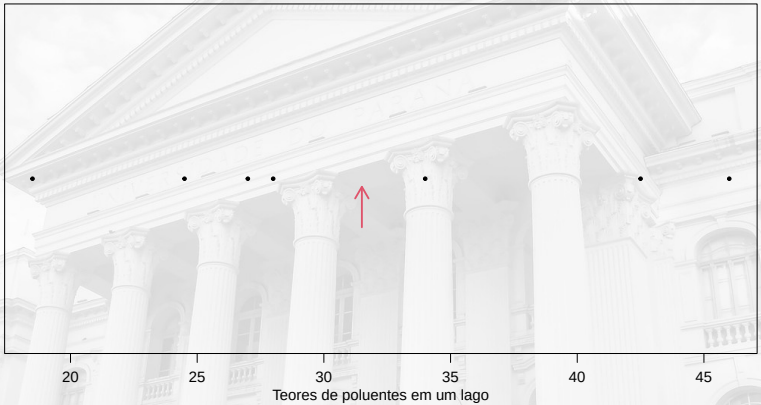
média: $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

um desvio individual: $e_i = y_i - \overline{y} \;\; (e_1 = 1, e_2 = -1, e_3 = -1, \ldots, e_{14} = -1, e_{15} = 0)$

variância: $S^2 = \dfrac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$

desvio padrão: $S = \sqrt{\dfrac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}}$

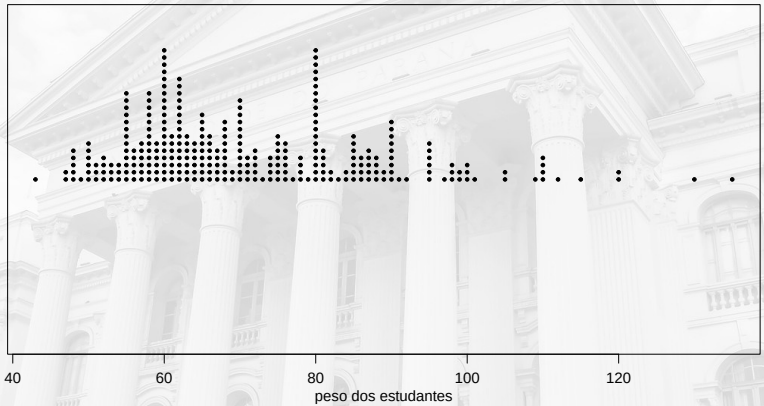
Uma interpretação da média e variância



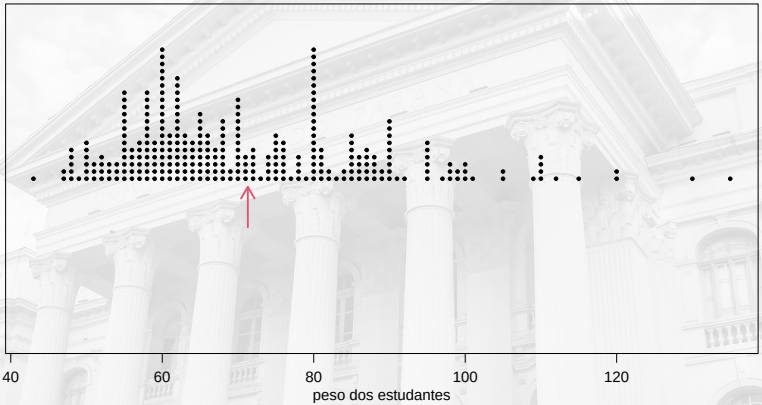
Média: Centro de massa ou de gravidade!

E a variância/desvio padrão?

Uma interpretação da média e variância



Uma interpretação da média e variância



Média: Centro de massa ou de gravidade!

E a variância/desvio padrão?

Exemplo 03: Número de artigos

Mais medidas de posição

$$2, 0, 0, 3, 0, 1, 1, 0, 0, 0, 1, 4, 2, 0, 1$$

Valor	0	1	2	3	4
Frequencia	7	4	2	1	1

a moda é 0.

Ordenando dados: $0, 0, 0, 0, 0, 0, 0, \mathbf{1}, 1, 1, 1, 2, 2, 3, 4$

Metade dos valores estão abaixo e metade acima da mediana **1**

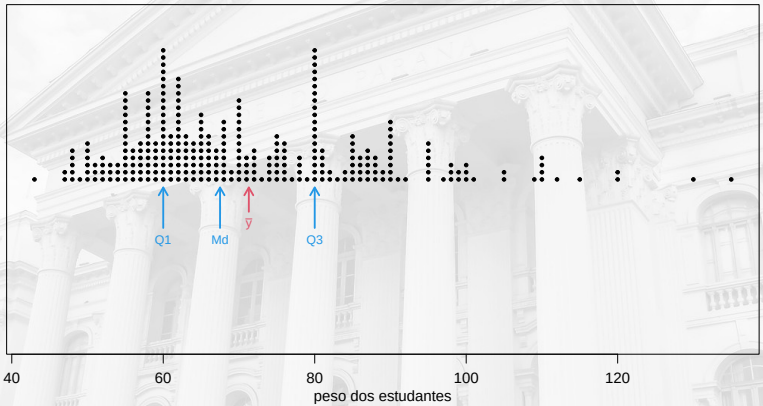
1/4 dos dados estão abaixo do 1^o quartil = 0
e 1/4 dos dados estão acima do 3^o quartil = 1,5

A mediana é o 2^o quartil.

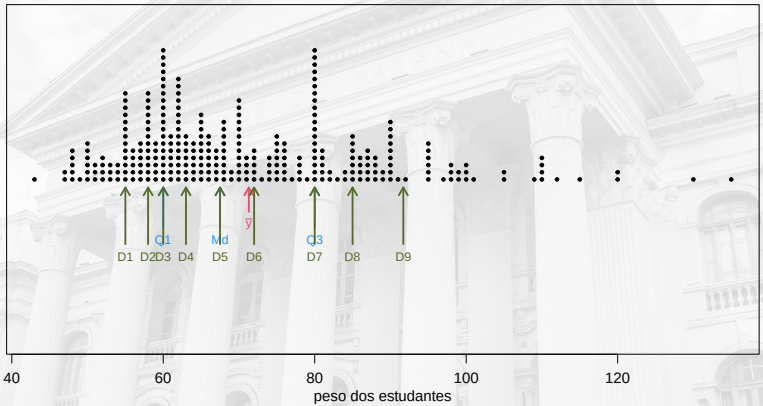
Os três quartis dividem os dados em quatro partes.

Os $(n_q - 1)$ quantis dividem os dados em n_q partes.

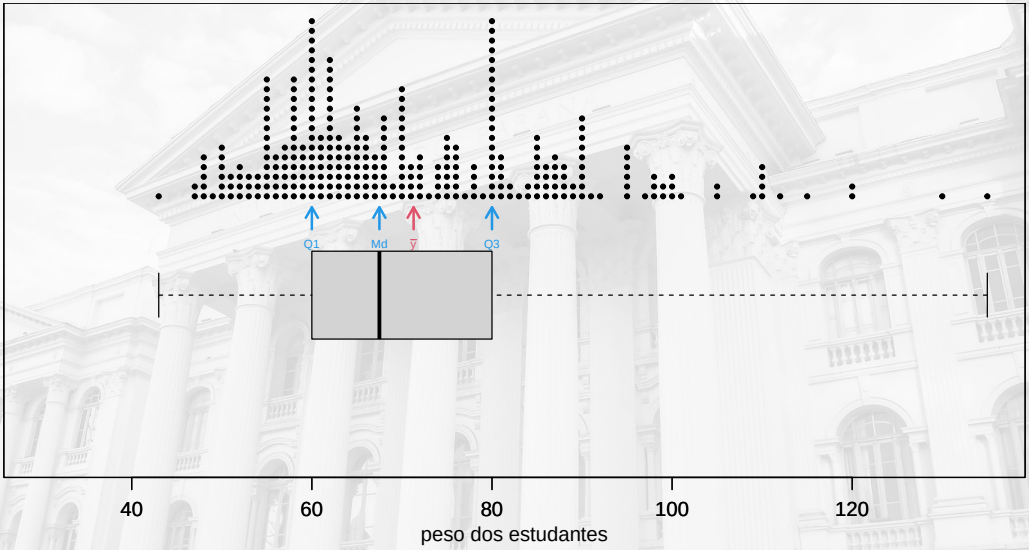
Mediana e quartis



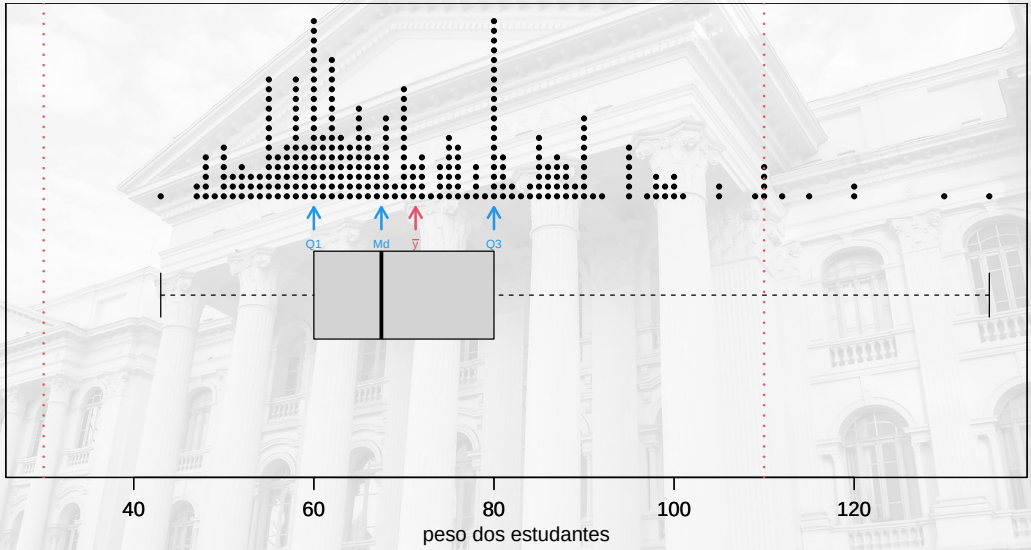
Decis


D1
D2
Q1
D3
D4
Md
D5
D6
ȳ
Q3
D7
D8
D9
40
60
80
100
120
peso dos estudantes

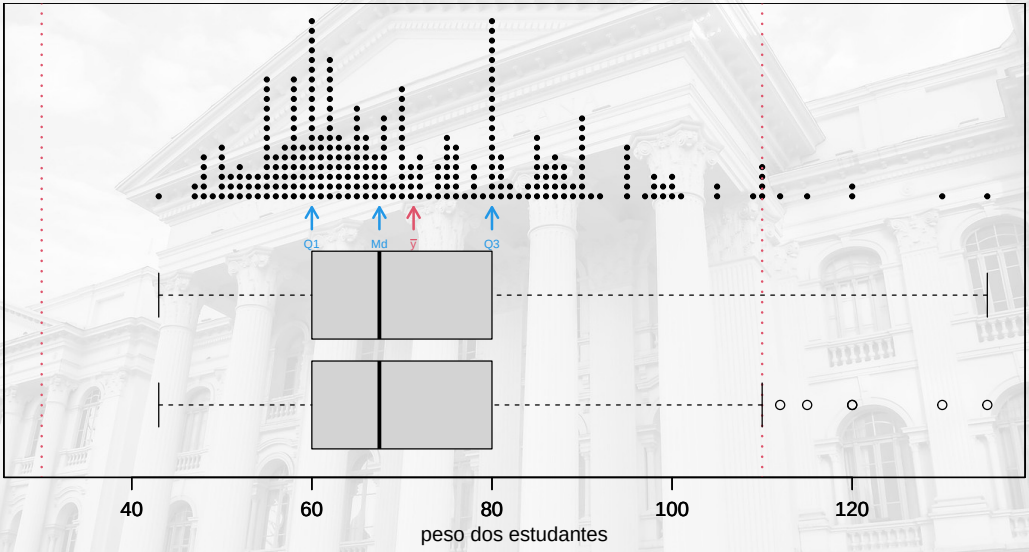
O Boxplot



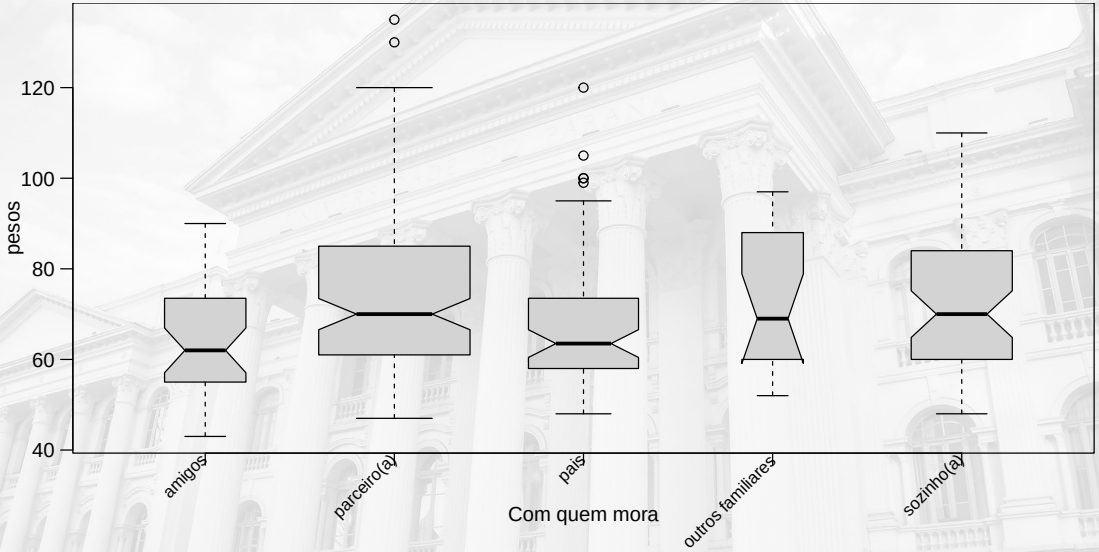
Boxplot



Boxplot



Variações no boxplot



Exemplo 03: Número de artigos

Mais medidas de dispersão

dados: $2, 0, 0, 3, 0, 1, 1, 0, 0, 0, 1, 4, 2, 0, 1$

dados ordenados: $0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 3, 4$

O menor dado é 0 e o maior é 4.

Portanto a amplitude é $A = y_n - y_1$ = 4 - 0 = 4.

O primeiro quantil é 0 e o terceiro é 1,5.

Portanto a amplitude interquartílica é $AI = Q_3 - Q_1$ = 1,5 - 0 = 1,5.

Comparando dispersões

A variabilidade é grande ou pequena?
Qual varia mais?

Exemplo	média	desvio padrão
teores no lago	31.5	9.9
número de artigos	1.0	1.3

Supondo outros conjuntos de dados:

Exemplo	média	desvio padrão
teores no lago (B)	63.0	9.9
número de artigos (B)	5.0	2.5

A variabilidade é a mesma?

Necessidade de uma medida de dispersão *relativa*.

Padronizando variabilidade

Medida de dispersão *relativa* ou *padronizada*

A variabilidade é grande ou pequena?

Exemplo	média	desvio padrão	C.V.
teores no lago	31.5	9.9	31.4 %
número de artigos	1.0	1.3	125.4 %

Coeficiente de variação (CV) $= 100 \times \frac{\text{desvio padrão}}{\text{média}}\%$.

- Medida adimensional
- Permite comparar variabilidade de grupos com médias diferentes
- Permite comparar variabilidade de variáveis/atributos diferentes

Padronizando dados

Medida de dispersão *relativa* ou *padronizada* dos dados

Quanto cada dado se afasta da média?

Exemplo do número de artigos: média = 1.0, d.p. = 1.3

dados	2	0	0	3	0	1	1	0	0	0	1	4	2	0	1
desvios	1	-1	-1	2	-1	0	0	-1	-1	-1	0	3	1	-1	0

Exemplo do lago: média = 31.5, d.p. = 9.9

dados	27	18,5	46	34	24,5	42,5	28
desvios	-4,5	-13,0	14,5	2,5	-7,0	11,0	-3,5

Padronizando dados

escore = $\frac{\text{dado - média}}{\text{desvio padrão}} \longrightarrow z_i = \frac{y_i - \overline{y}}{S_y}$

Exemplo: média = 1.0, d.p. = 1.3

dados	2	0	0	3	0	1	1	0	0	0	1	4	2	0	1
desvios	1	-1	-1	2	-1	0	0	-1	-1	-1	0	3	1	-1	0
escores	0,8	-0,8	-0,8	1,6	-0,8	0,0	0,0	-0,8	-0,8	-0,8	0,0	2,4	0,8	-0,8	0,0

Exemplo do lago: média = 31.5, d.p. = 9.9

dados	27	18,5	46	34	24,5	42,5	28
desvios	-4,5	-13,0	14,5	2,5	-7,0	11,0	-3,5
escores	-0,45	-1,31	1,46	0,25	-0,71	1,11	-0,35

Padronizando dados

Escores:

- São dados adimensionais.
- A média dos scores é 0 e a variância é 1.
- Permitem comparar dados de grupos com médias e/ou variabilidades diferentes.
- Permitem comparar dados de variáveis/atributos diferentes.

Notação

Expressar idéias e conceitos de forma mais clara, objetiva e sintética.

Y :	definição da variável
y :	valores de dados tomados da variavel
y_i :	o i-ésimo dado
$y =$	$\{y_1, y_2, y_3, \ldots y_n\}$
(y) :	dados ordenados
$(y) =$	$\{y_{(1)}, y_{(2)}, y_{(3)}, \ldots y_{(n)}\}$
$(y_{(1)}, y_{(n)})$:	mínimo e máximo
$\overline{Y}$:	a média da variável
$\overline{y}$:	a média dos dados
$\mathrm{md}(Y)$:	a mediana da variável
$\mathrm{md}(y)$:	a mediana dos dados

Notação e Fórmulas

Y : Teor do elemento no lago

$$y = \{y_1 = 27, y_2 = 18{,}5, y_3 = 46, y_4 = 34, y_5 = 24{,}5, y_6 = 42{,}5, y_7 = 28\}$$

$$(y) = \{y_{(1)} = 18{,}5, y_{(2)} = 24{,}5, y_{(3)} = 27, y_{(4)} = 28, y_{(5)} = 34, y_{(6)} = 42{,}5, y_{(7)} = 46\}$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 31.5$$

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1} = 98.0$$

$$S_y = \sqrt{S_y^2} = 9.9$$

$$\text{CV}_y = \frac{S_y}{\overline{y}} = 31.4$$

$$\text{A}_y = y_{(n)} - y_{(1)} = 27{,}5$$

$$\text{Q}_{1y} = 25{,}75$$

$$\text{Q}_{2y} = \text{med}(y) = 28$$

$$\text{Q}_{3y} = 38{,}25$$

$$\text{AI}_y = Q_3 - Q_1 = 12{,}5$$

Médias ...

Diferentes tipos de médias.

- Aritmética: $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.
- Ponderada: $\overline{y} = \frac{\sum_{i=1}^{n} w_i \cdot y_i}{\sum_{i=1}^{n} w_i}$.
- Geométrica: $\overline{y} = \left(\prod_{i=1}^{n} y_i\right)^{1/n} = \sqrt[n]{y_1 \cdot y_2 \ldots y_n} = \exp\left(\frac{\sum_{i=1}^{n} \ln y_i}{n}\right) = \exp\left(\overline{\ln y_i}\right)$.
- Harmônica: $\overline{y} = \frac{n}{\sum_{i=1}^{n} 1/y_i} = \left(\frac{\sum_{i=1}^{n} y_i^{-1}}{n}\right)^{-1} = \left(\overline{y^{-1}}\right)^{-1}$.
- Aparada (*trimmed*),
- ...
- Como resultado de uma *otimização*.
- ...

Um exemplo

Foi tomada uma amostra medindo-se o teor de um certo elemento contaminante na água em diferentes pontos de um lago. Deseja-se avaliar a contaminação do lago e ainda decidir se uma intervenção deve ser feita se houver evidências de que o teor ultrapassa 38 un.

Dados obtidos (amostra):

27 18,5 46 34 24,5 42,5 28

Como estimar o teor no lago como um todo?
Onde foram coletados os dados

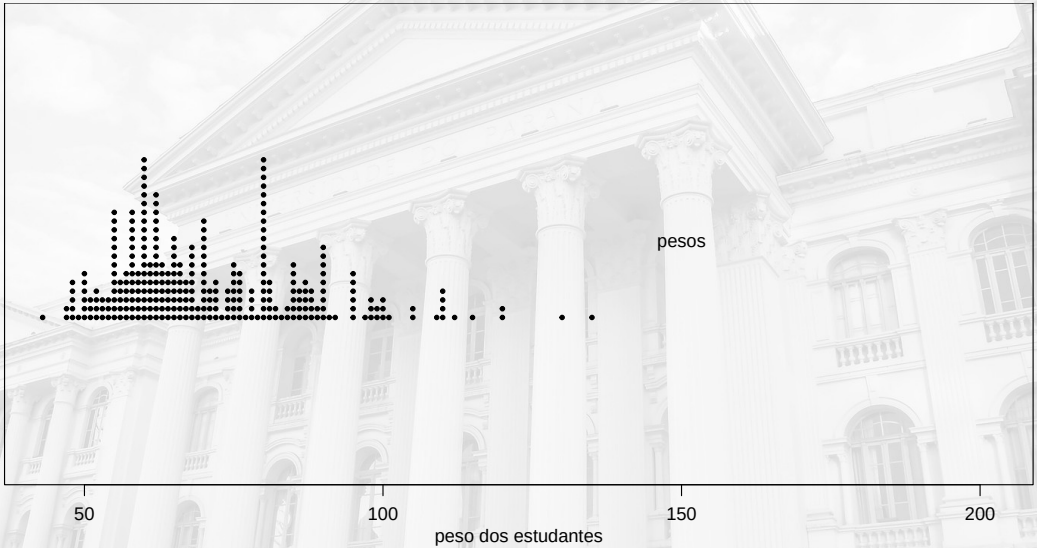
Um exercício

Uma empresa tem 200 funcionários e uma folha de pagamento de R$ 600.000,00. Os salários pagos variam desde R$ 900,00 até R$ 23.000,00. A empresa decidiu dar um aumento para seus funcionários e está para decidir entre duas formas de aumento:

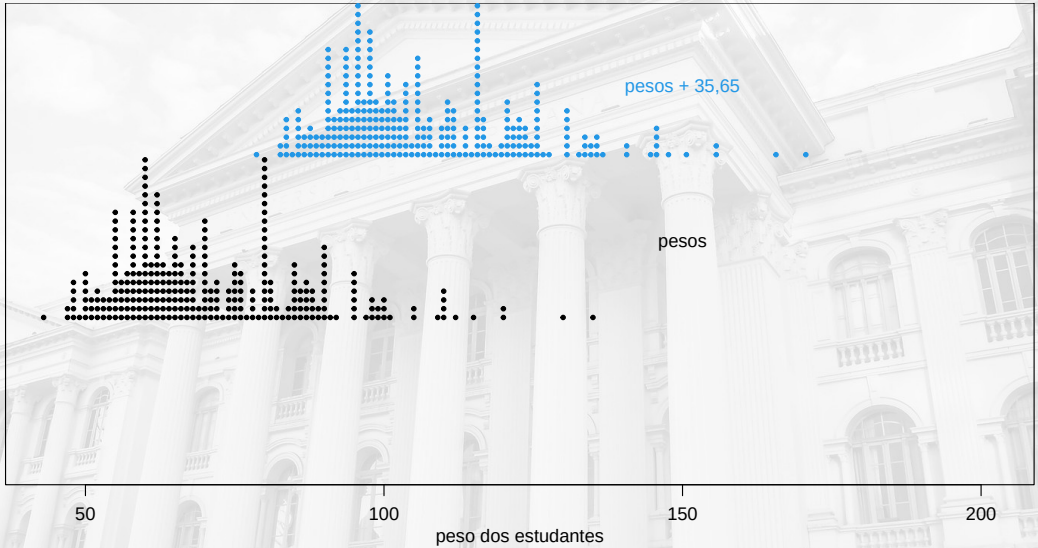
1. dar um aumento de 10% para todos os funcionários;
2. dar um aumento de R$ 300,00 para todos os funcionários.

Discuta o(s) impacto(s) de cada proposta e se há diferenças entre elas.

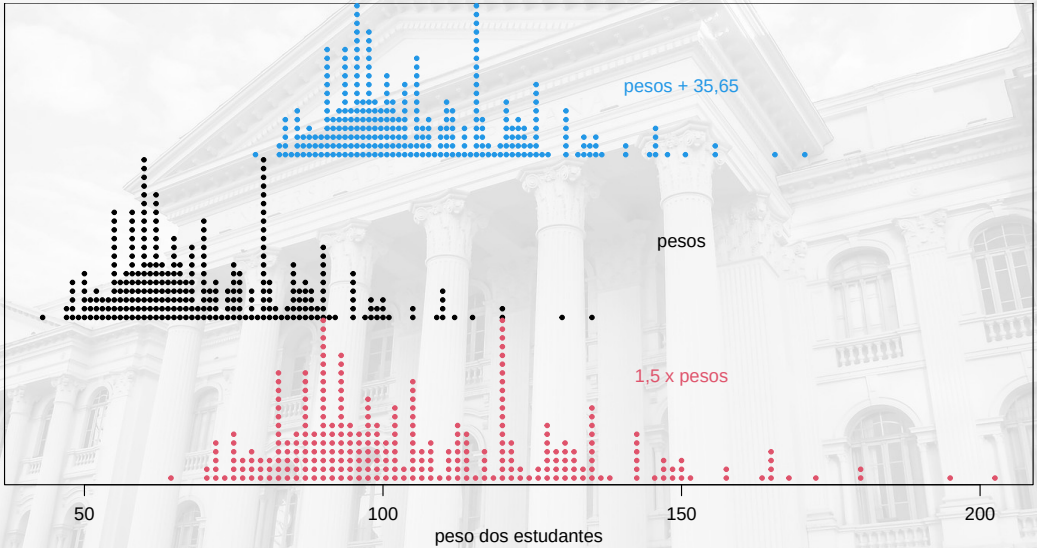
Ilustrando com outros dados



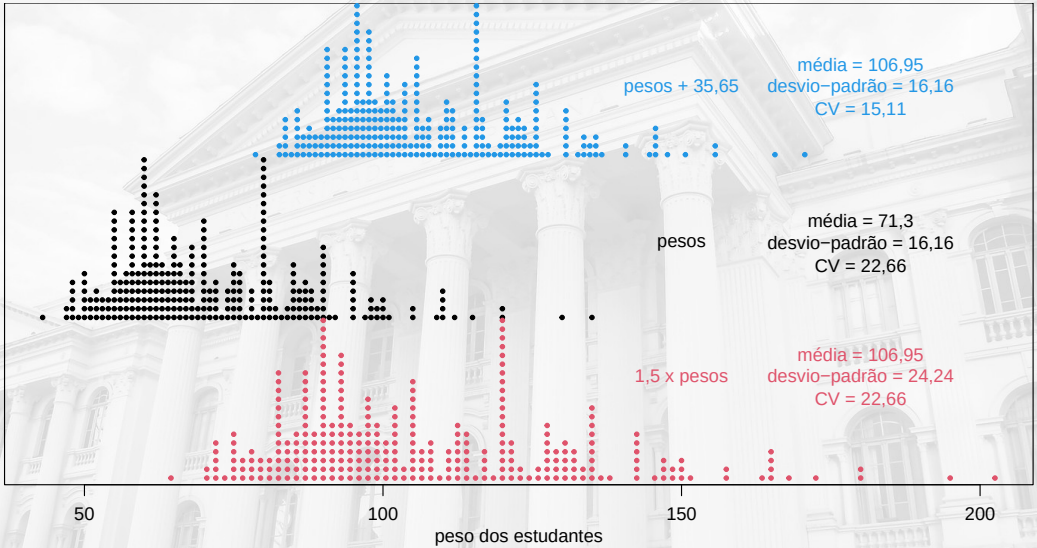
Ilustrando com outros dados



Ilustrando com outros dados



Ilustrando com outros dados



Formalizando

Y : salários $\qquad Y_1 = Y + k$: salários com aumento de valor constante $k > 0$

$$\overline{Y} = \frac{\sum Y_i}{n} \qquad \overline{Y_1} = \frac{\sum Y_{1i}}{n} = \frac{\sum Y_i + k}{n} = \frac{\sum Y_i}{n} + \frac{kn}{n} = \overline{Y} + k$$

$$S_y^2 = \frac{\sum (y_i - \overline{y})^2}{n-1} \qquad S_{y_1}^2 = \frac{\sum (y_{1i} - \overline{y}_1)^2}{n-1} = \frac{\sum [(y_i + k) - (\overline{y} + k)]^2}{n-1} = \frac{\sum (y_i - \overline{y})^2}{n-1} = S_y^2$$

$$S_y = \sqrt{S_y^2} \qquad S_{y_1} = S_y$$

$$CV_y = 100 \times \frac{S_y}{\overline{Y}} \qquad CV_{y_1} = 100 \times \frac{S_{y_1}}{\overline{Y}_{y_1}} = 100 \times \frac{S_y}{\overline{Y} + k} < CV_y$$

Formalizando

Y : salários

$Y_2 = kY$: salários com aumento percentual

$$\overline{Y} = \frac{\sum Y_i}{n}$$

$$\overline{Y_2} = \frac{\sum Y_{2i}}{n} = \frac{\sum kY_i}{n} = k\frac{\sum Y_i}{n} = k\overline{Y}$$

$$S_y^2 = \frac{\sum (y_i - \overline{y})^2}{n-1}$$

$$S_{y_2}^2 = \frac{\sum (y_{2i} - \overline{y}_2)^2}{n-1} = \frac{\sum (ky_i - k\overline{y})^2}{n-1} = k^2 \frac{\sum [(y_i - \overline{y})^2}{n-1} = k^2 S_y^2$$

$$S_y = \sqrt{S_y^2}$$

$$S_{y_2} = kS_y$$

$$CV_y = 100 \times \frac{S_y}{\overline{Y}}$$

$$CV_{y_2} = 100 \times \frac{S_{y_2}}{Y_{y_2}} = 100 \times \frac{kS_y}{k\overline{Y}} = 100 \times \frac{S_y}{\overline{Y}} = CV_y$$

Resumindo com notação

Traduza em palavras ...

$$\overline{Y} \pm k \qquad \overline{Y \pm k} = \overline{kY} = k\overline{Y}$$

$$S_{y \pm k} = S_y \qquad S_{ky} = kS_y$$

$$CV_{y \pm k} \lessgtr CV_y \qquad CV_{ky} = CV_y$$

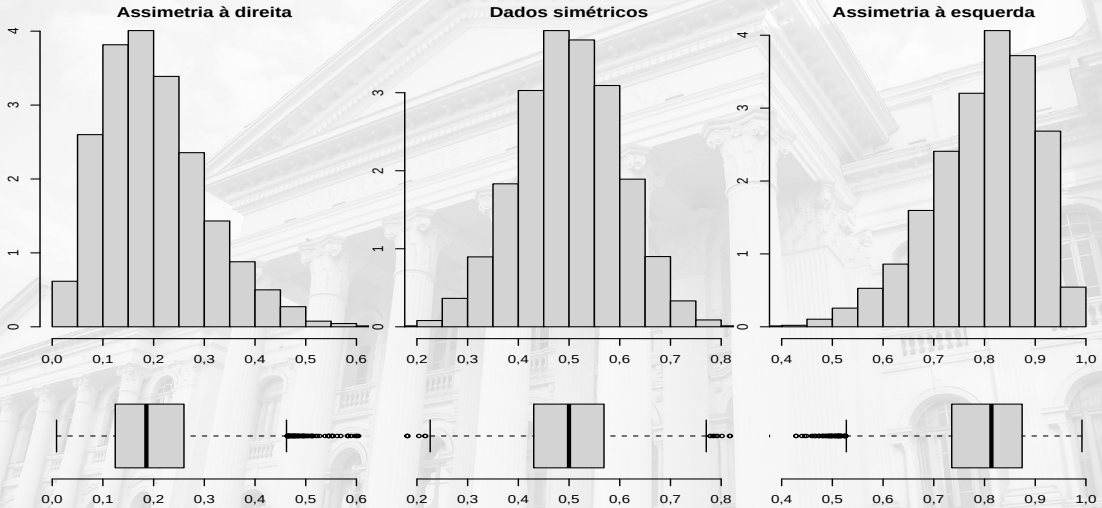
...e interpretamos no contexto do exemplo.

Revisitando um slide

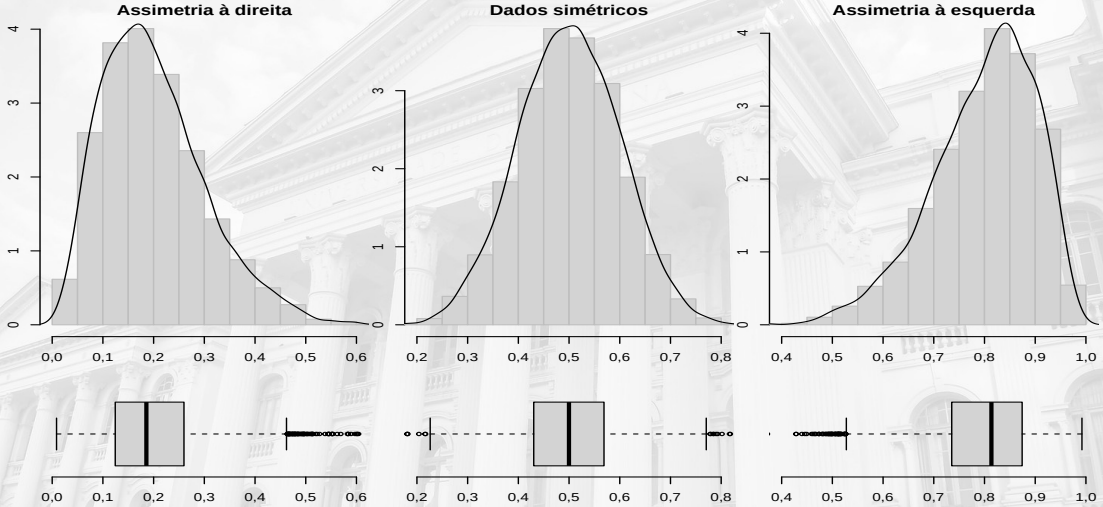
Resumindo variável numérica: Intuição

- *Por onde andam* os dados? (posição)
- Variam muito ou pouco? (dispersão)
- Como são distribuídos?
- Existem dados atípicos?
- Melhor expressos em outra escala? (transformação)

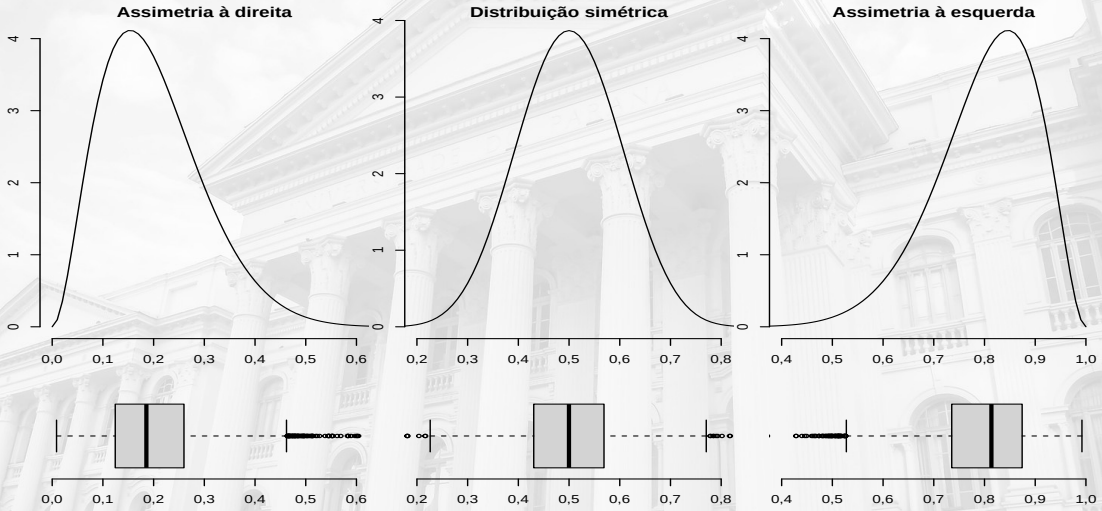
Comportamentos estilizados de variáveis contínuas



Comportamentos estilizados de variáveis contínuas

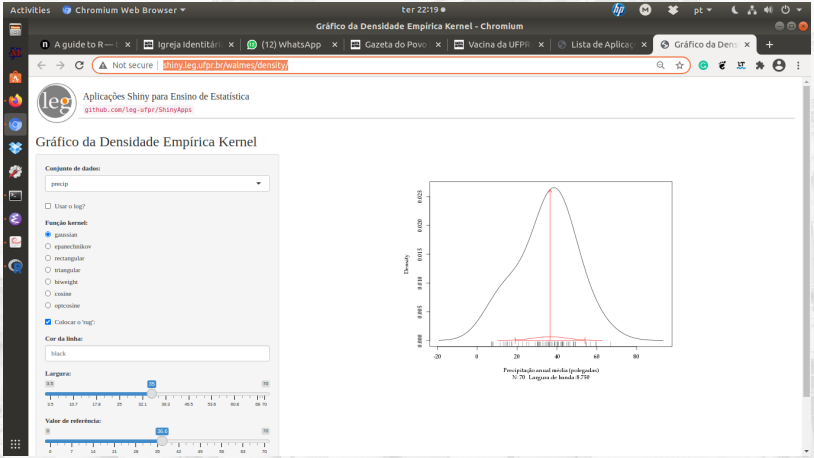


Comportamentos estilizados de variáveis contínuas



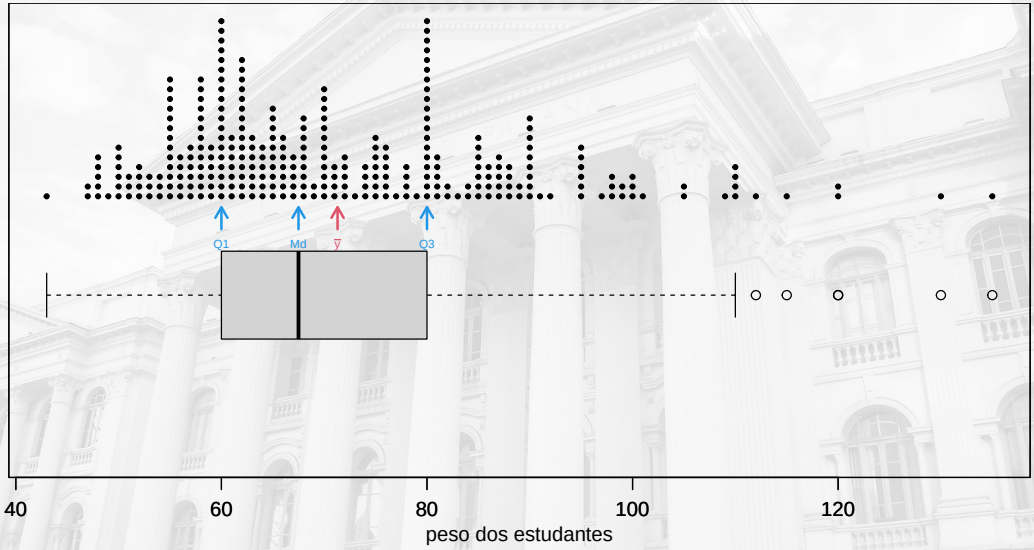
O gráfico de estimação de densidade

Disponível em: (http://shiny.leg.ufpr.br/walmes/density)

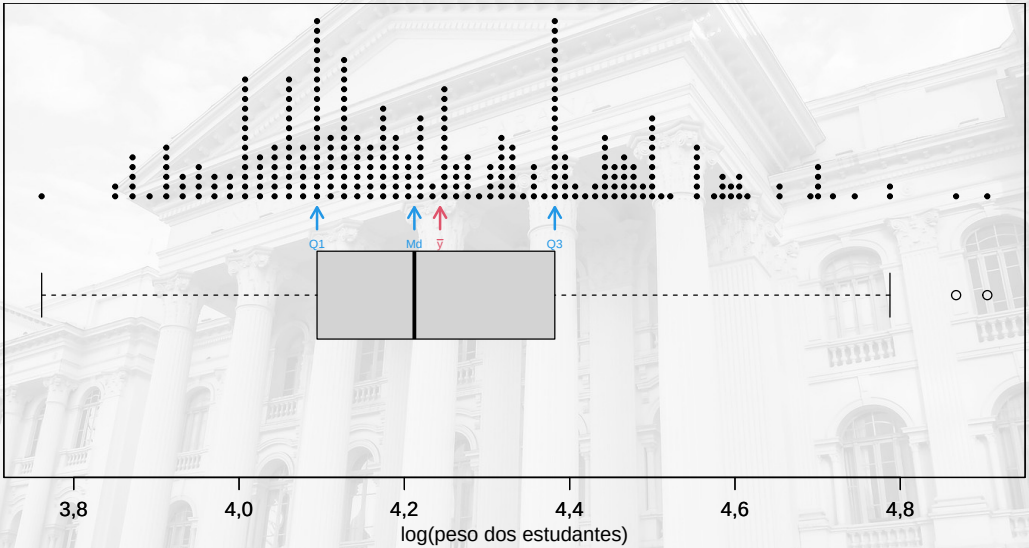


Autor: Prof. Walmes Zeviani

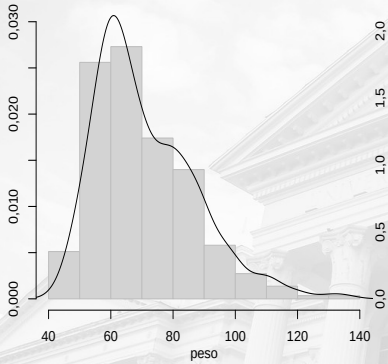
Revisitando pesos

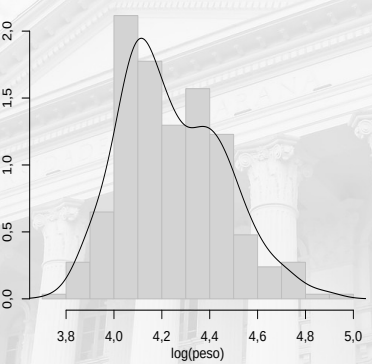


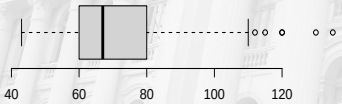
Transformando: log(pesos)

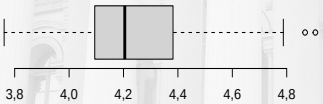


Transformando: log(pesos)









- Qual o comportamento estilizado?
- Escala logarítmica ou original?
- Por que log?
- Outra transformação possível?
- bimodal?

Transformações ("normalizadoras")

Qual adotar?

- $\log(y)$
- $\sqrt{y}$
- $1/y$
- $1/\sqrt{y}$
- $\log(\frac{y}{1-y})$
- $\text{seno}(y)$
- ...

- Opção pelo contexto.
- Opção pela natureza do dado.
- Tentativa e erro
- Busca da *melhor*
  Famílias de transformação:
  - transformação de Box-Cox,
  - outras famílias de transformações.

Transformação Box-Cox

$$Y^t = \begin{cases} \frac{Y^\lambda - 1}{\lambda} \text{ se } \lambda \leq 0 \\ \log(y) \text{ se } \lambda = 0 \end{cases}$$

Encontrar λ define uma transformação "adequada". (Tipicamente $\lambda \in [-2,2]$)

Alguns casos particulares:

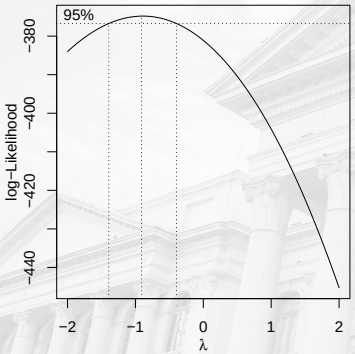
- $\lambda = 0 \rightarrow \log(y)$
- $\lambda = 0{,}5 \rightarrow \sqrt{y}$
- $\lambda = -1 \rightarrow 1/y$
- $\lambda = -0{,}5 \rightarrow 1/\sqrt{y}$

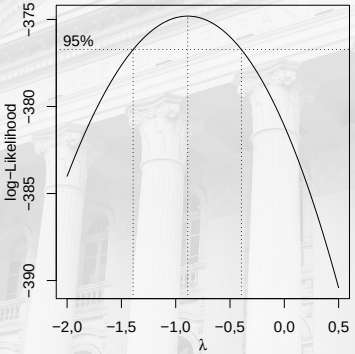
Outras transformações:

- $\log(\frac{y}{1-y})$
- $\text{seno}(y)$
- $\text{argtanh}(y)$
- ...

Transformação Box-Cox

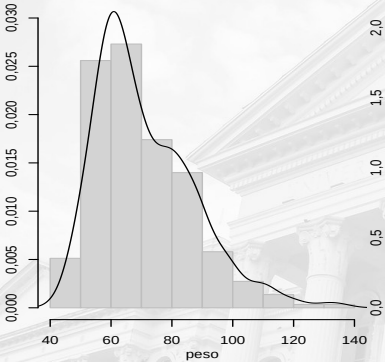
Determinando λ:

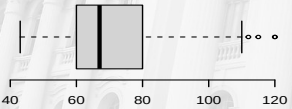


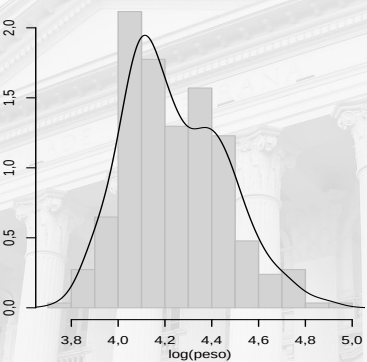


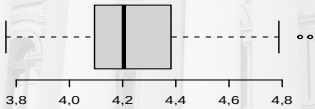
- Gráfico original em $[-2, 2]$.
- *Zoom* do gráfico em $[-2, 0.5]$.
- Melhor estimativa $\lambda = -0.89$.
- $\lambda \approx -1$.
- Estimativa aceitável e interpretável.
- $y^* = 1/y$

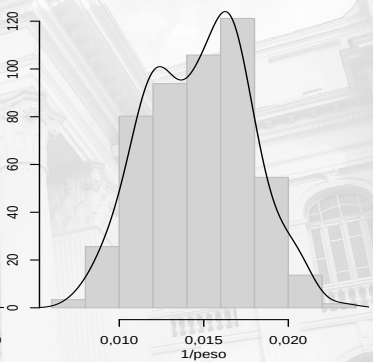
Dados de pesos e transformações

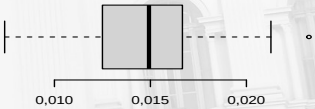




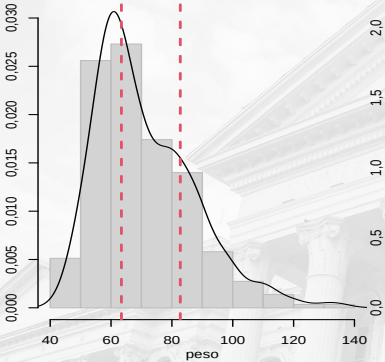


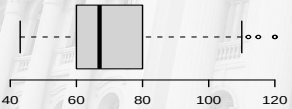


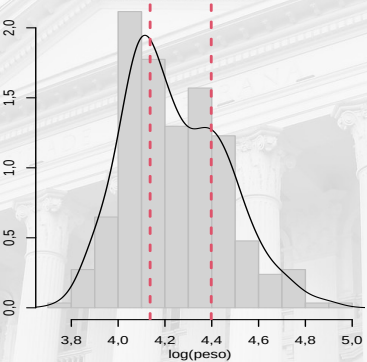


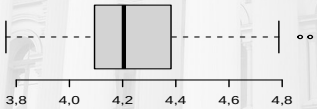


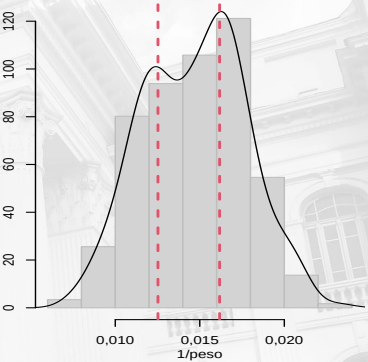
Dados de pesos e transformações

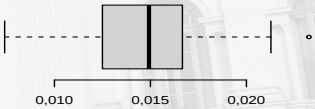




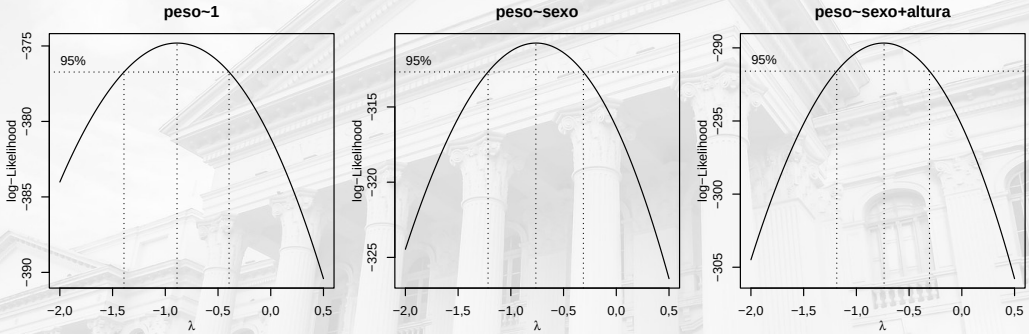








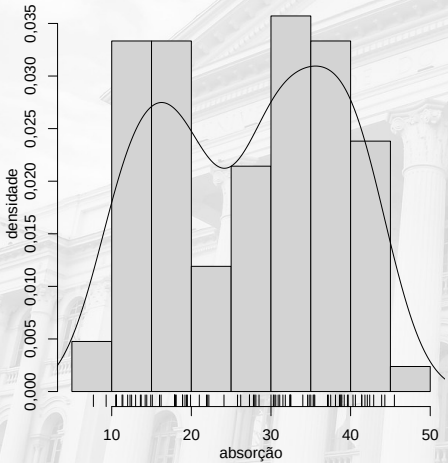
Box-Cox com covariáveis

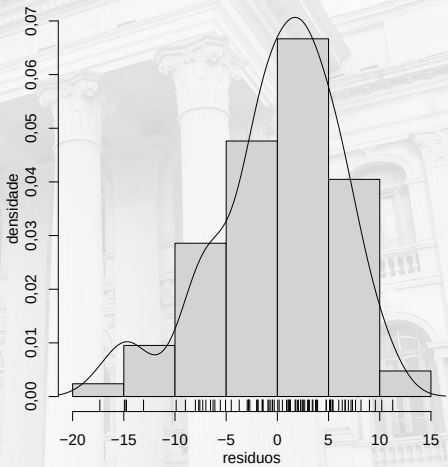


Neste exemplo as covariáveis não influiram na escolha transformação.
Mas isto nem sempre é o caso.

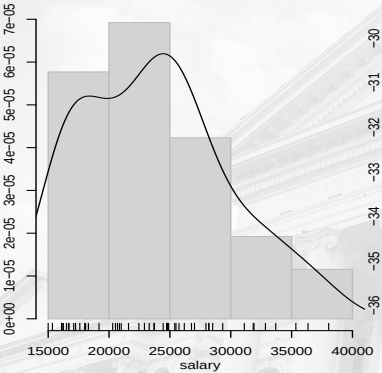
Transformação Box-Cox: cuidado!

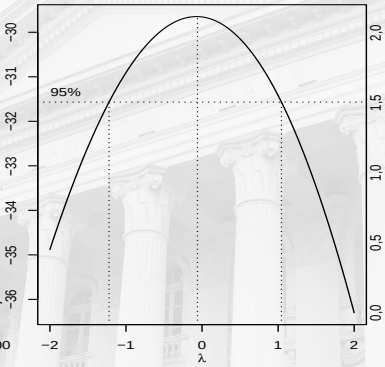
Dados experimentais: absorção de carbono para tipos de plantas e condições.
Cuidado com respostas! "Descontar efeito de preditores".

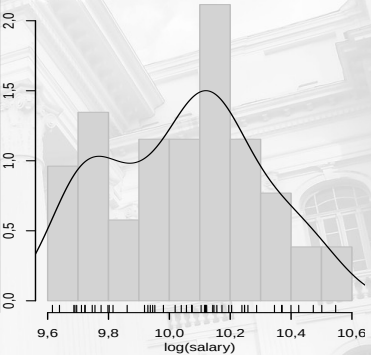


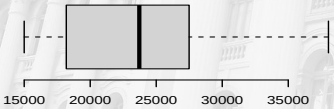


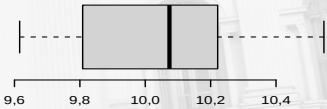
Outro exemplo: dados de salários





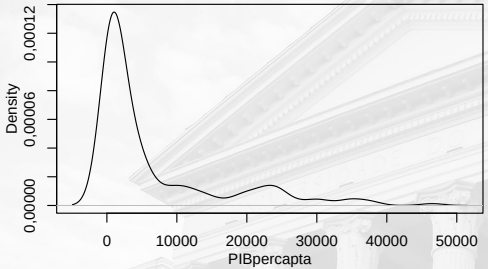


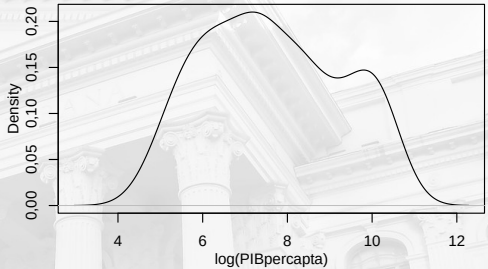


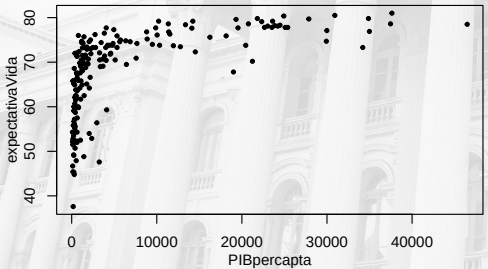


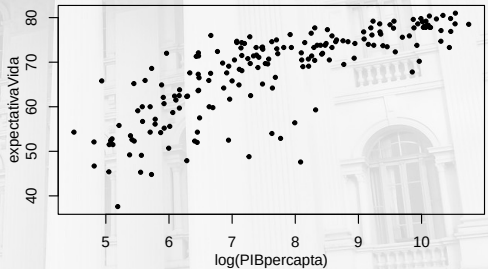
1

Exemplo: dados gapminger - Ano 2000









Referências bibliográficas

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

MAGALHÃES, M. N.; LIMA, A. C. P. de. **Noções de Probabilidade e Estatística**. 7. ed. São Paulo: Edusp, 2015.

UTTS, J. M. **Seeing Through Statistics**. [S.l.: s.n.], 2005.

WILD, C. J.; SEBER, G. A. F. **Chance Encounters: A First Course in Data Analysis and Inference**. [S.l.: s.n.], 2000.

WILD, C. J.; SEBER, G. A. F. **Encontros Com O Acaso. Primeiro Curso De Análise De Dados e Inferência**. [S.l.: s.n.], 2004.

ZEVIANI, W. et al. EstBas: Um curso em estatística básica. <http://www.leg.ufpr.br/estbas>. 2021.