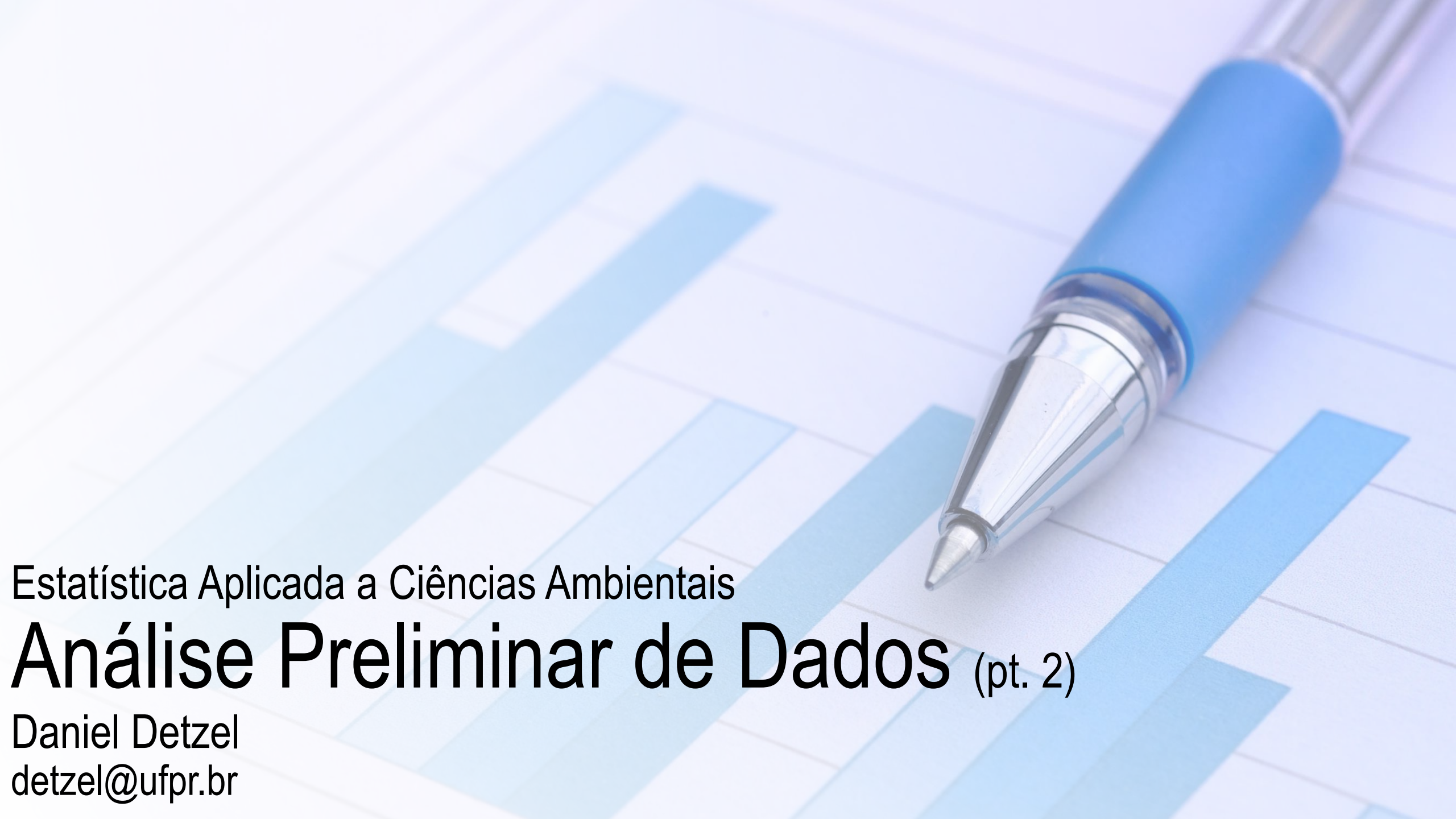


Estatística Aplicada a Ciências Ambientais

Análise Preliminar de Dados (pt. 2)

Daniel Detzel
detzel@ufpr.br

Agenda

Representação gráfica de dados

- cuidados na representação gráfica

Apresentação dos trabalhos da semana

Estatísticas descritivas

- tendência central
- variabilidade
- assimetria

Análise Preliminar de Dados

cuidados na representação gráfica

Análise preliminar de dados | cuidados na representação gráfica

"ao menos 65% das pessoas são aprendizes visuais"
(Dr. Richard Felder, em tradução livre)

Quando as análises estatísticas são finalizadas, é hora de apresentá-las

Para apresentações orais

figuras são (muito!) mais eficientes que tabelas

Para apresentações escritas

figuras precisam ser autoexplicativa (com ajuda de uma legenda apropriada)

Análise preliminar de dados | cuidados na representação gráfica

Recomendações gerais para um bom gráfico

- cuide com cores: maximize a relação “informação/tinta”
- evite perspectivas
- eixos devem começar em zero quando a magnitude do dado é importante
- conexões entre pontos somente com variáveis contínuas
- apresentações orais devem conter gráficos diferentes de apresentações por escrito

Análise preliminar de dados | cores

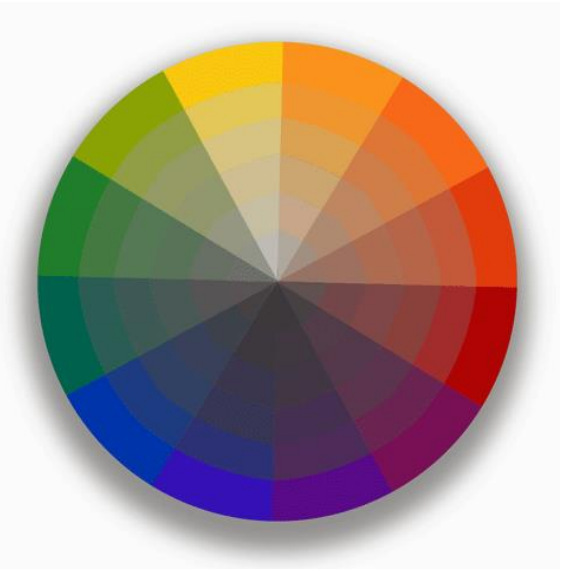
Cuidados com cores

podem aumentar o interesse do público, porém enviesar a interpretação de resultados

Características trazidas pelas cores

cores quentes ou intensas (grande saturação) tendem a aumentar os objetos

cores frias ou em tons pastéis amenizam o efeito



Fonte: https://medium.theuxblog.com/basics-of-color-theory-83abc55ce94b

Análise preliminar de dados | cores

Características trazidas pelas cores (cont.)

dados representados por círculos/pontos com diferentes cores são mais eficientes do que representados por diferentes símbolos

linhas com diferentes cores são mais eficientes do que linhas sólidas vs. linhas tracejadas (ou outros padrões)

Cuidado com múltiplas cores

pessoas conseguem diferenciar entre 8 a 12 cores

daltonismo (mais comum: distinção entre vermelho e verde)

sugestão: https://www.color-blindness.com/coblis-color-blindness-simulator/

Análise preliminar de dados | cores

Relação “informação/tinta” (*data-ink ratio*) (Tufte, 1983)

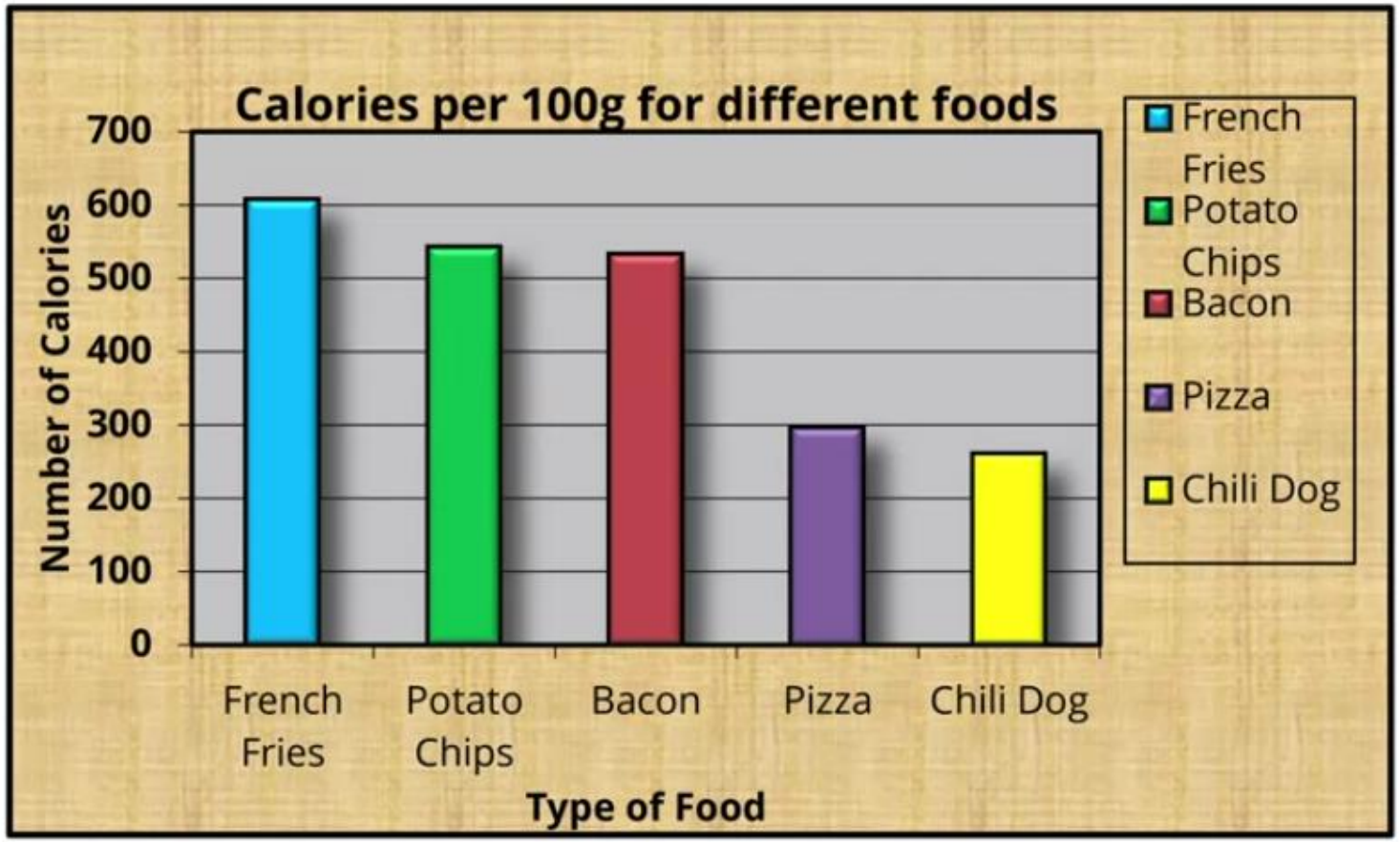
$$\textit{Data-ink ratio} = \frac{\text{cores relevantes}}{\text{total de cores do gráfico}}$$

Em essência: somente os objetos que contém as informações relevantes devem ser destacados por cores

Tufte, E. The Visual Display of Quantitative Information, 1983.

Análise preliminar de dados | cores

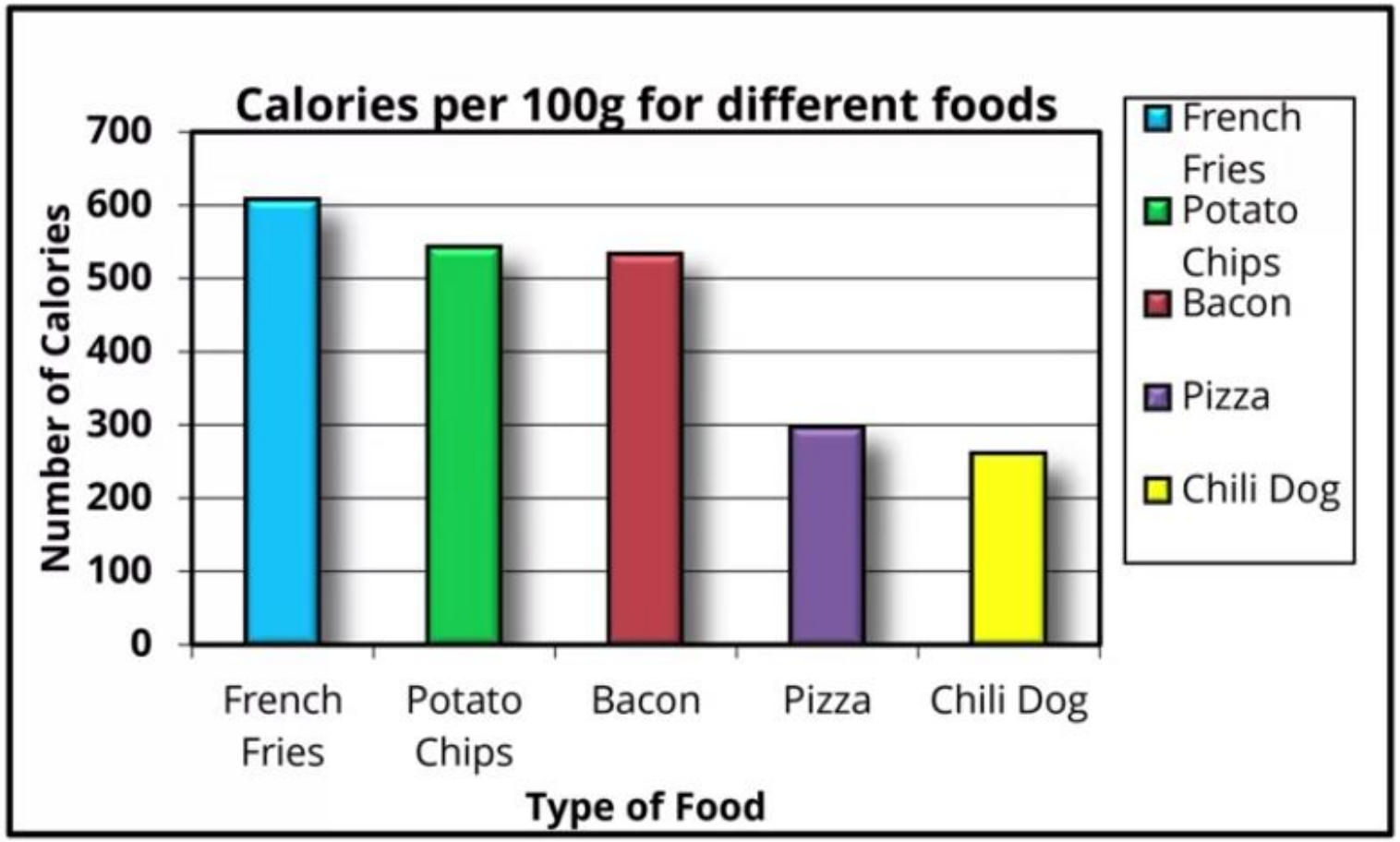
Exemplo de maximização da *data-ink ratio*: calorias em 100g de bacon



Fonte: https://www.codeconquest.com/blog/data-ink-ratio-explained-with-example/

Análise preliminar de dados | cores

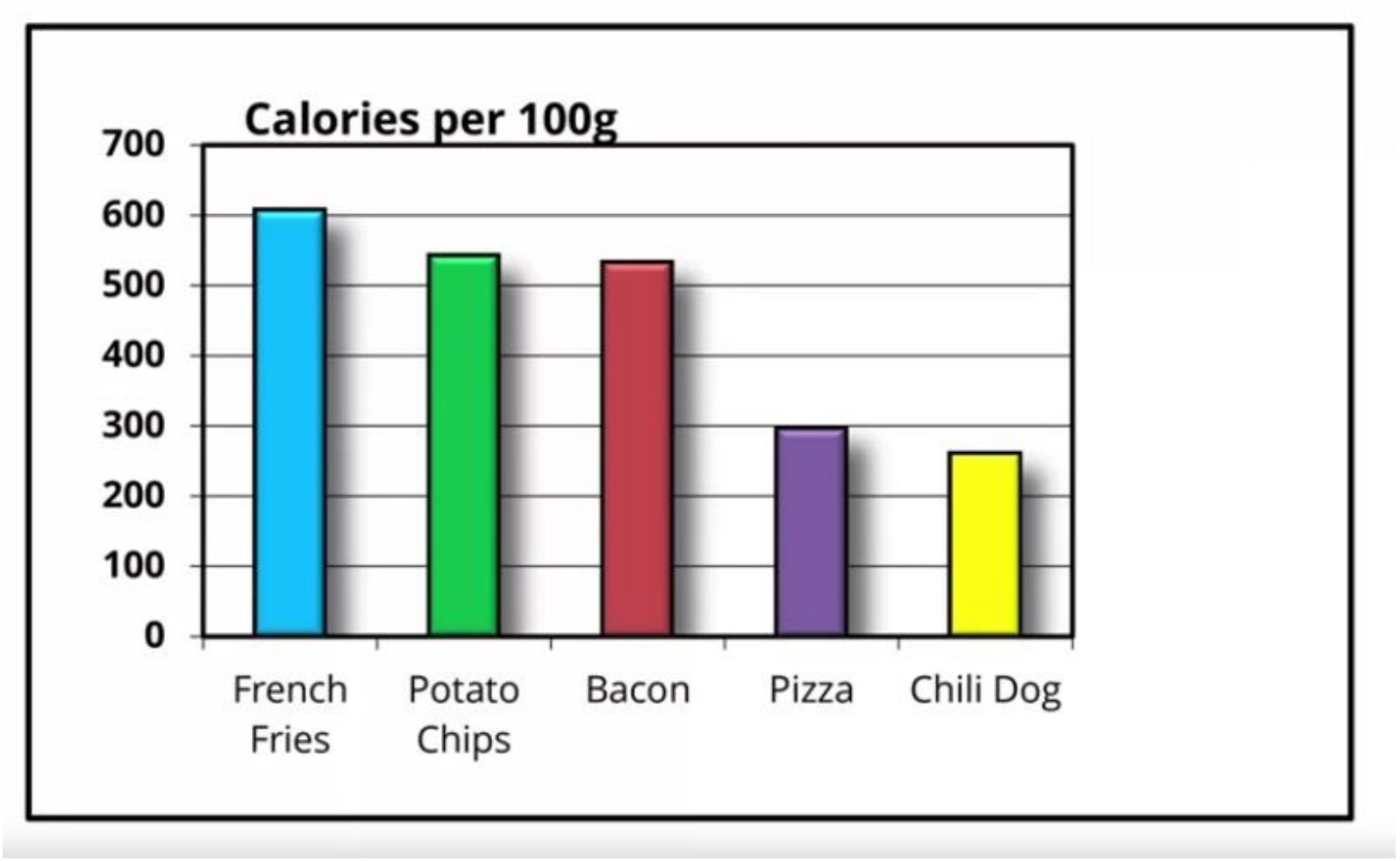
1. Eliminação das cores de fundo



Fonte: https://www.codeconquest.com/blog/data-ink-ratio-explained-with-example/

Análise preliminar de dados | cores

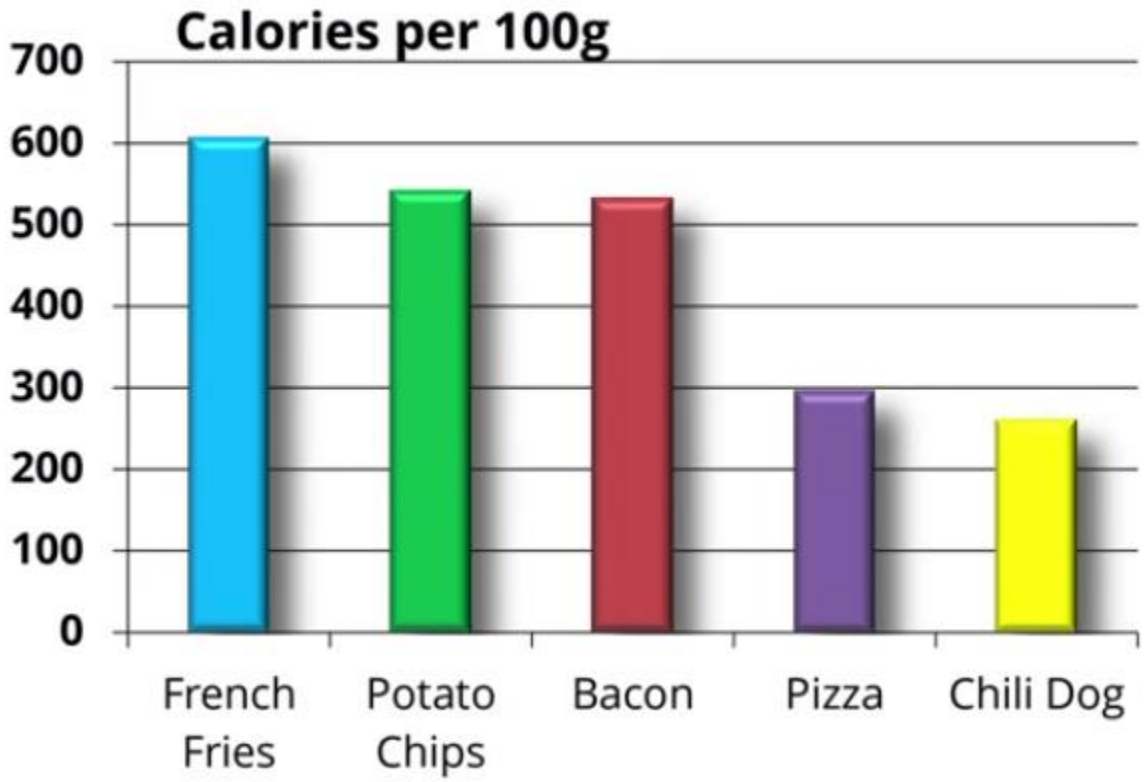
2. Eliminação de informações redundantes (legenda e títulos)



Fonte: https://www.codeconquest.com/blog/data-ink-ratio-explained-with-example/

Análise preliminar de dados | cores

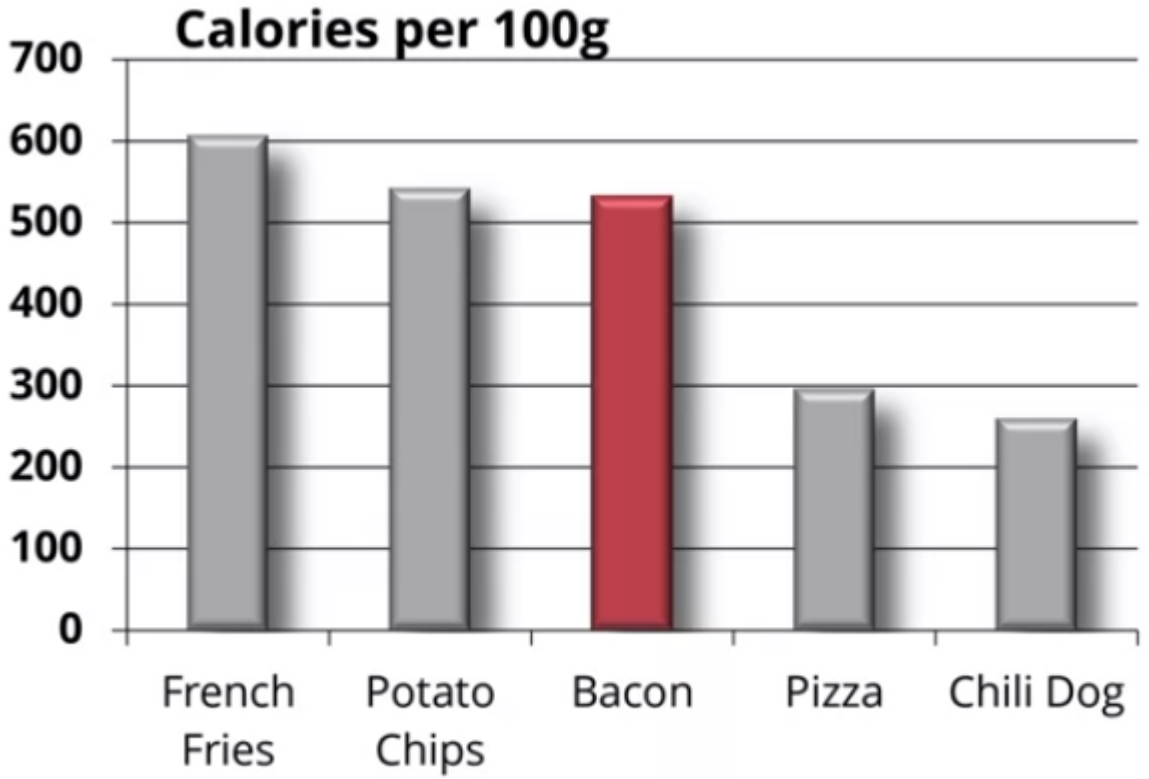
3. Eliminação das bordas



Fonte: https://www.codeconquest.com/blog/data-ink-ratio-explained-with-example/

Análise preliminar de dados | cores

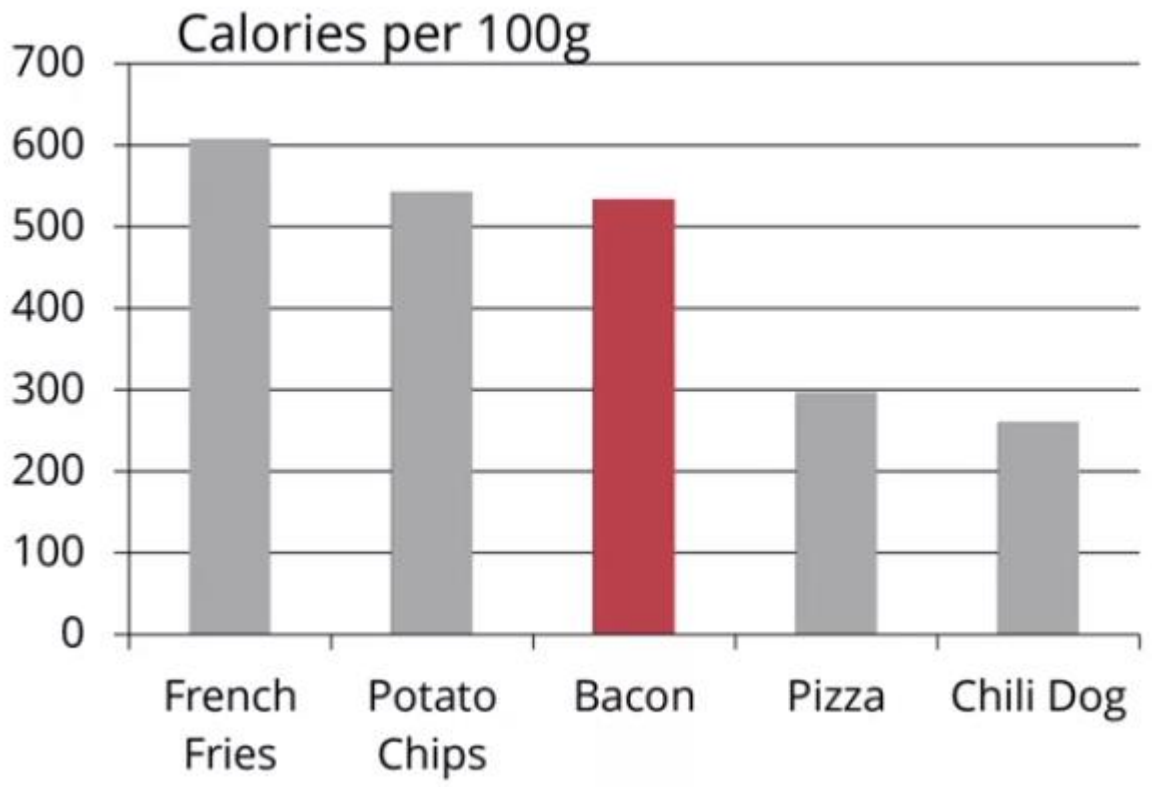
4. Aplicação de cores somente à informação de destaque



Fonte: https://www.codeconquest.com/blog/data-ink-ratio-explained-with-example/

Análise preliminar de dados | cores

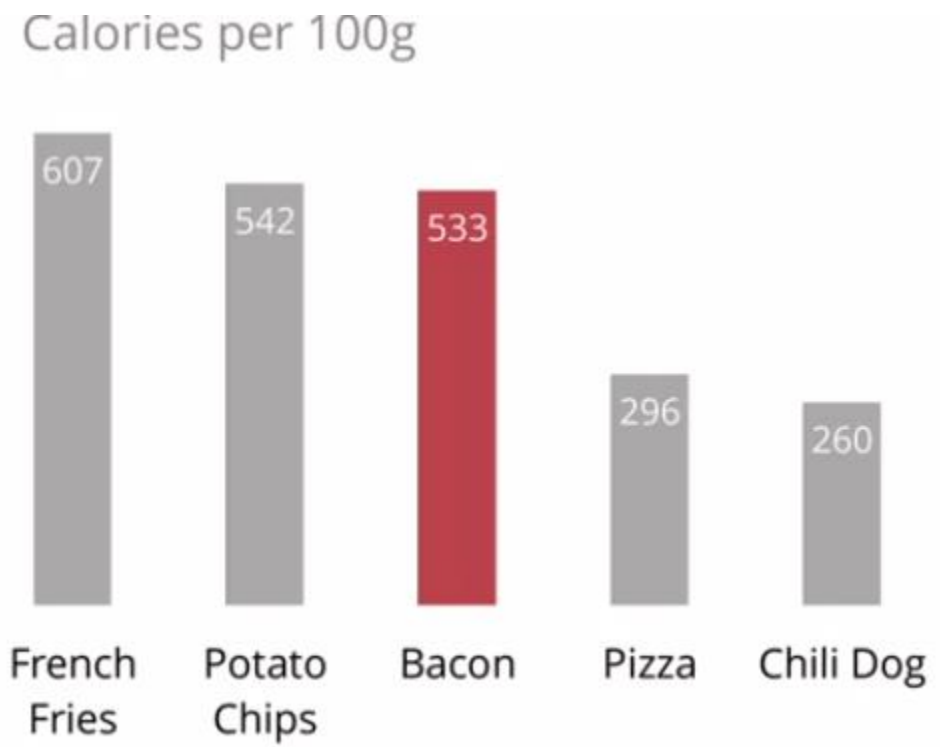
5. Retirada sombras e perspectivas



Fonte: https://www.codeconquest.com/blog/data-ink-ratio-explained-with-example/

Análise preliminar de dados | cores

6. Eliminação de linhas de grade

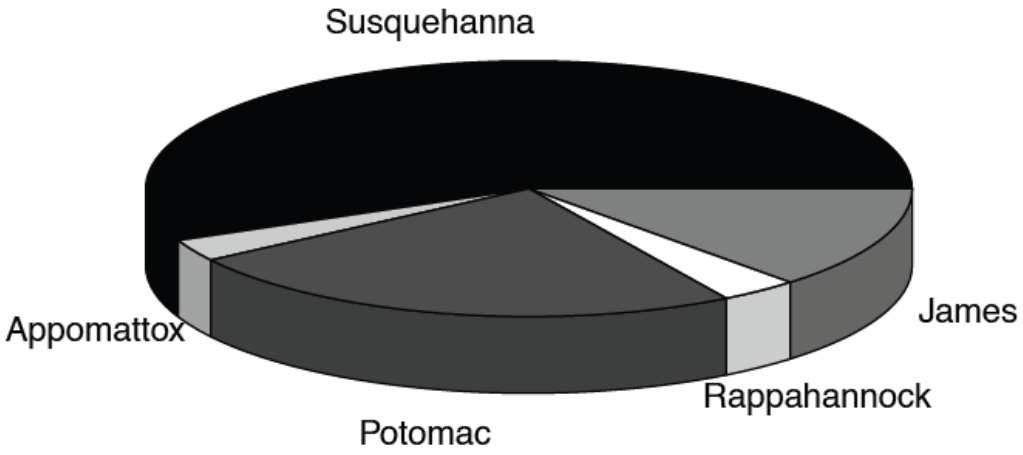


Fonte: https://www.codeconquest.com/blog/data-ink-ratio-explained-with-example/

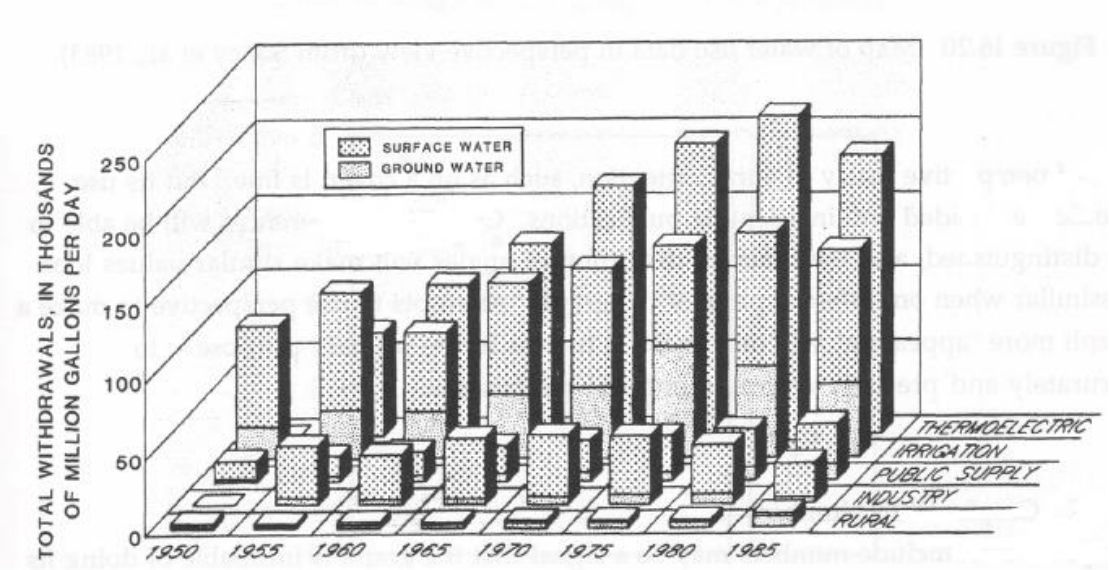
Análise preliminar de dados | perspectiva

Perspectiva: aspecto de 3 dimensões

- prejudica análises que envolvem áreas, comprimentos ou ângulos
- nosso cérebro compensa a perspectiva aumentando o tamanho de objetos à distância



Área de drenagem de 5 bacias na Baía de Chesapeake (EUA)
Fonte: Helsel et al. (2020, p. 423)



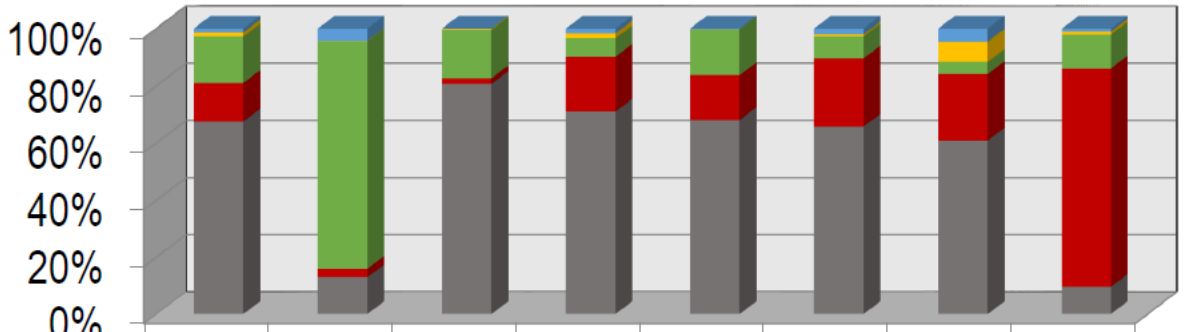
Usos da água nos EUA
Fonte: Helsel et al. (2020, p. 423)

Análise preliminar de dados | gráficos com números

Gráficos com números (além dos eixos)

podem indicar que o gráfico precisa ser mais bem definido

quando requeridos, números devem ser mostrados em outro lugar (apêndices)



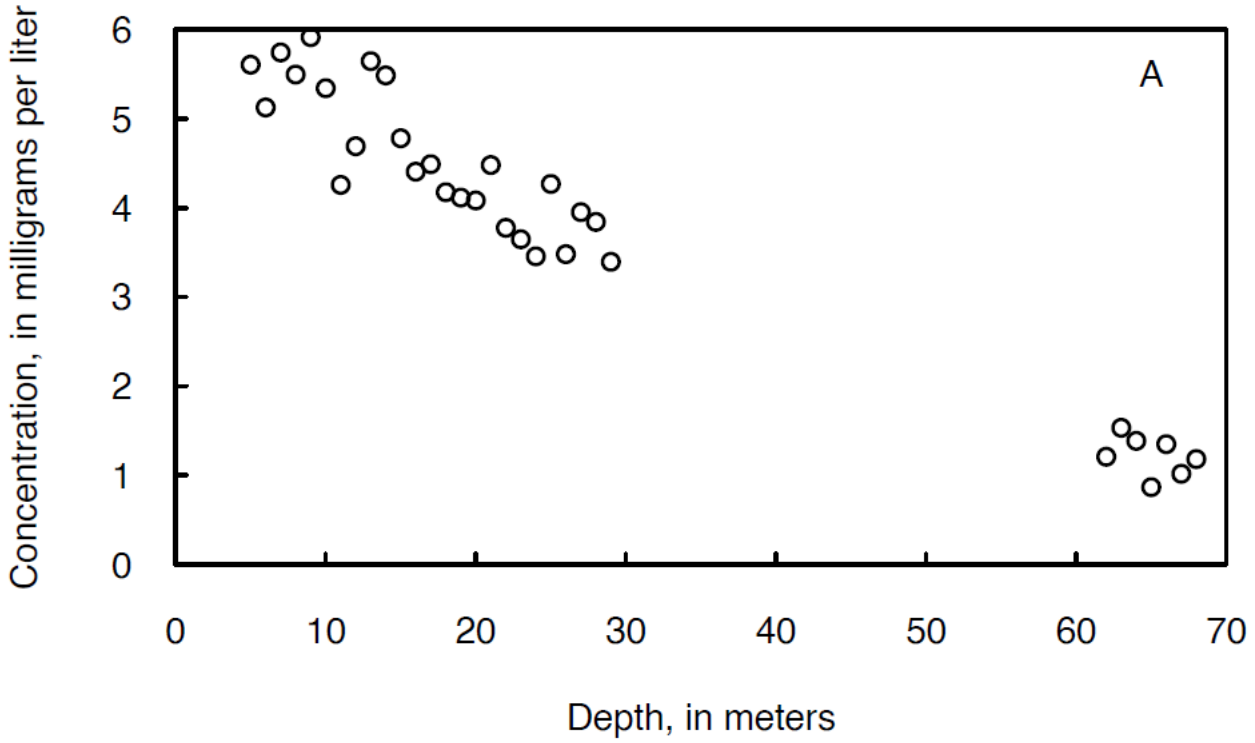
	Mundo	BRA	CHI	EUA	RUS	JAP	ALE	FRA
Biomassa	1,34%	4,30%	0,07%	1,67%	0,24%	2,04%	4,55%	1,03%
Eólica/Solar	1,47%	0,13%	0,38%	1,73%	0,05%	0,71%	7,12%	1,09%
Hidro	16,23%	79,64%	16,93%	6,45%	16,04%	7,68%	4,23%	11,85%
Nuclear	13,48%	3,01%	1,97%	19,17%	15,66%	23,87%	23,21%	76,49%
Gás	21,23%	6,24%	0,90%	20,84%	47,55%	26,19%	13,80%	3,83%
Óleo	5,48%	3,87%	0,67%	1,33%	1,54%	12,86%	1,44%	1,01%
Carvão	40,78%	2,80%	79,09%	48,80%	18,92%	26,65%	45,64%	4,70%

Análise preliminar de dados | escalas

Quebras de escala

cuidar para não confundir o público

preferir uma quebra completa na escala



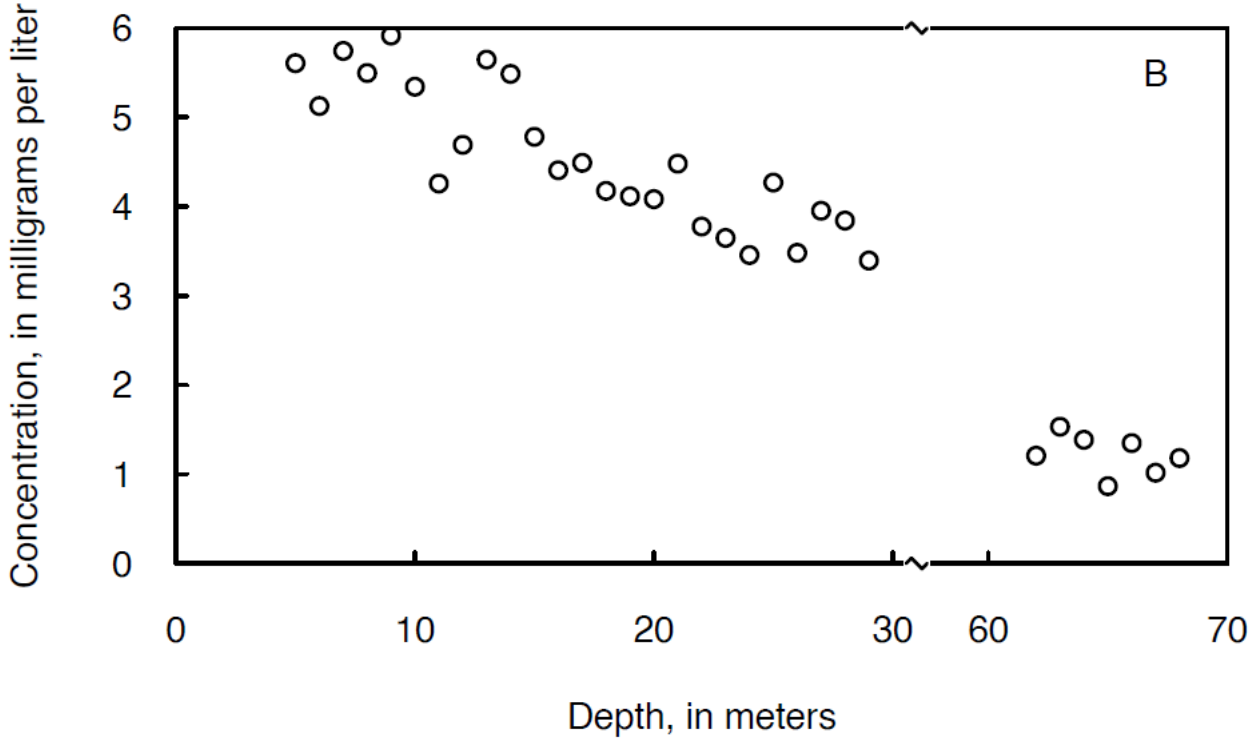
Fonte: Helsel et al. (2020, p. 426)

Análise preliminar de dados | escalas

Quebras de escala

cuidar para não confundir o público

preferir uma quebra completa na escala



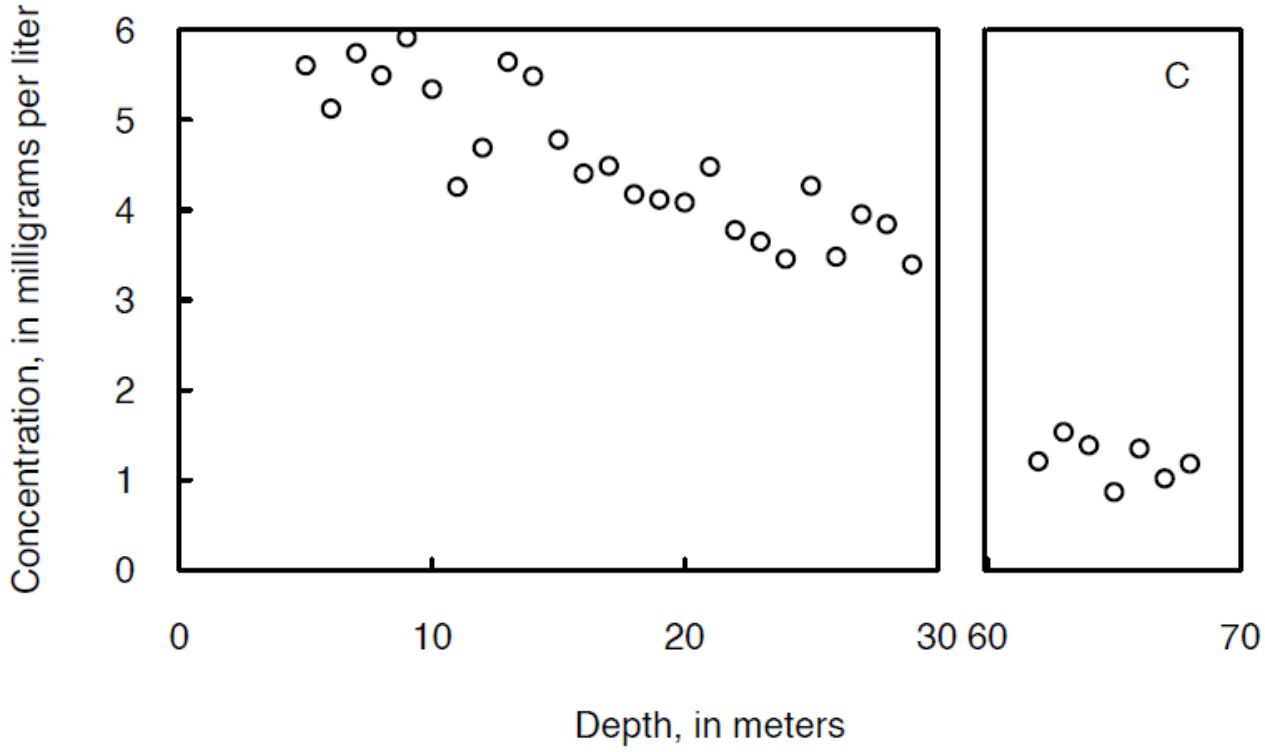
Fonte: Helsel et al. (2020, p. 426)

Análise preliminar de dados | escalas

Quebras de escala

cuidar para não confundir o público

preferir uma quebra completa na escala



Fonte: Helsel et al. (2020, p. 426)

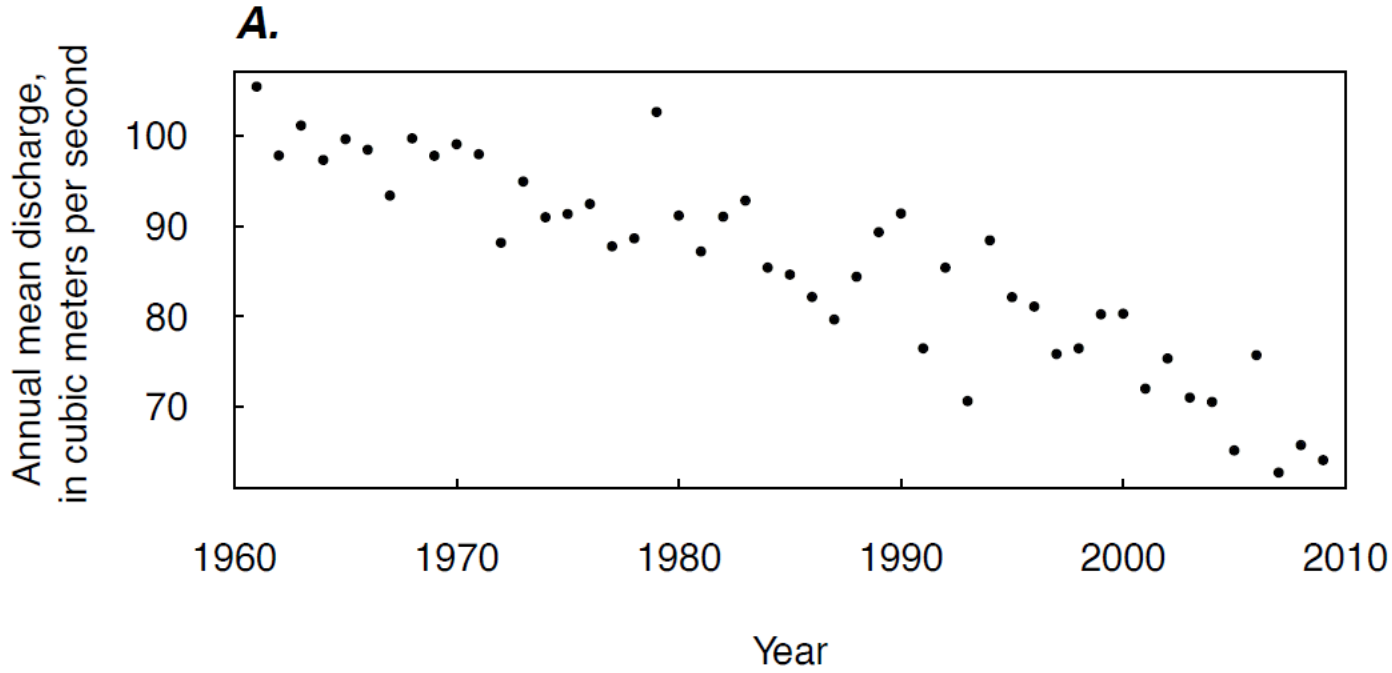
Análise preliminar de dados | escalas

Escalas automáticas

ferramentas ajustam automaticamente as escalas dos eixos

podem não ser apropriadas para a apresentação das conclusões desejadas

ex.: análise de tendências



Fonte: Helsel et al. (2020, p. 427)

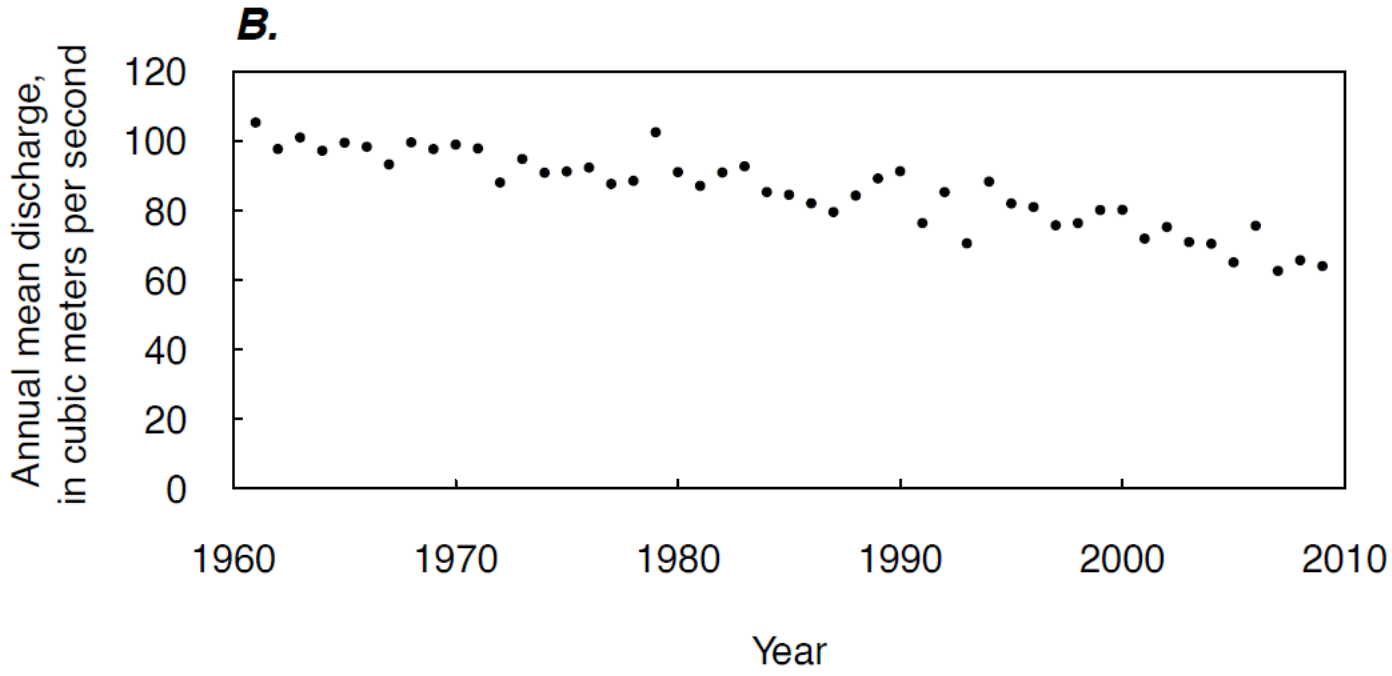
Análise preliminar de dados | escalas

Escalas automáticas

ferramentas ajustam automaticamente as escalas dos eixos

podem não ser apropriadas para a apresentação das conclusões desejadas

ex.: análise de tendências

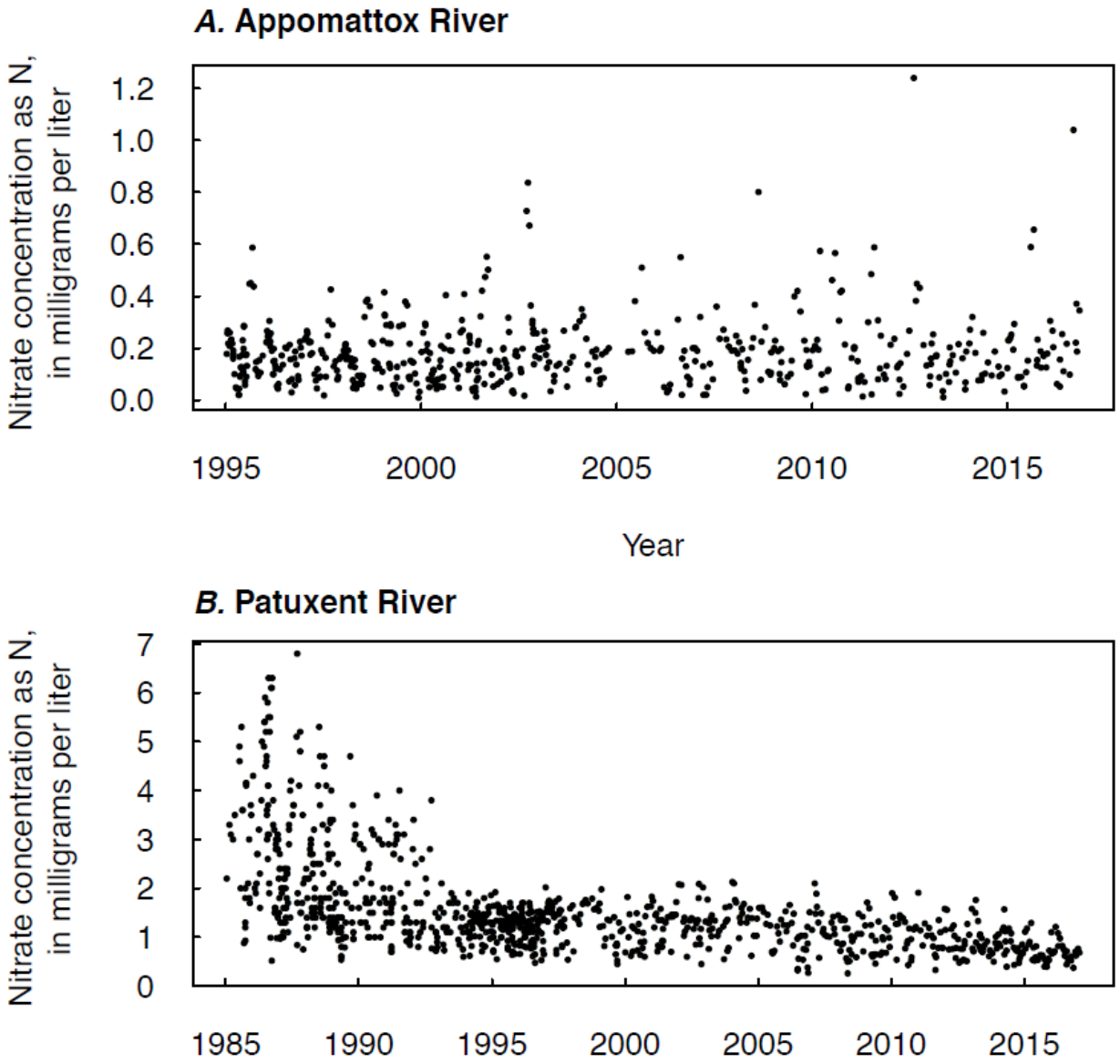


Fonte: Helsel et al. (2020, p. 427)

Análise preliminar de dados | escalas

Escalas automáticas

ex.: comparação entre dados

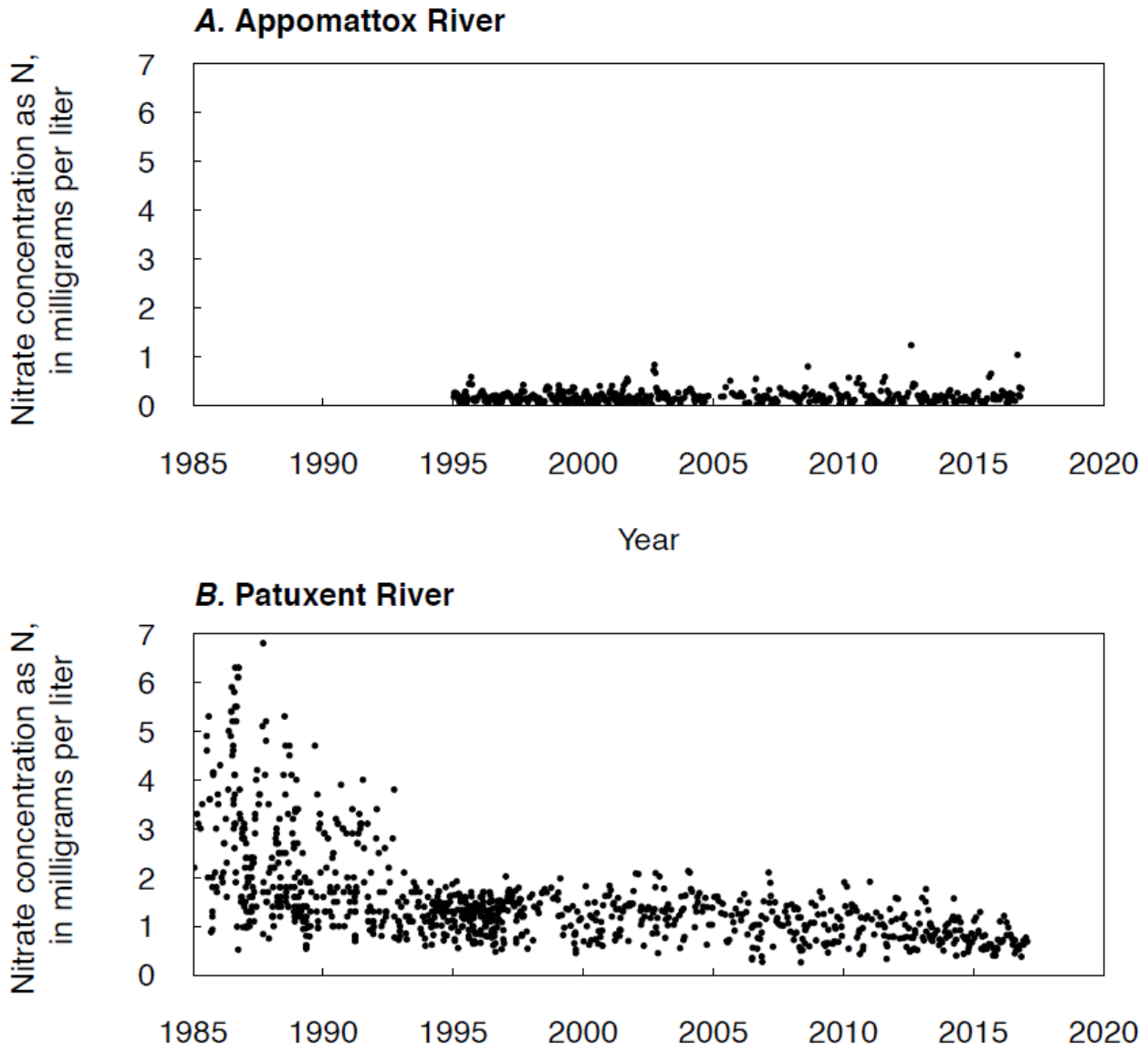


Fonte: Helsel et al. (2020, p. 428)

Análise preliminar de dados | escalas

Escalas automáticas

ex.: comparação entre dados



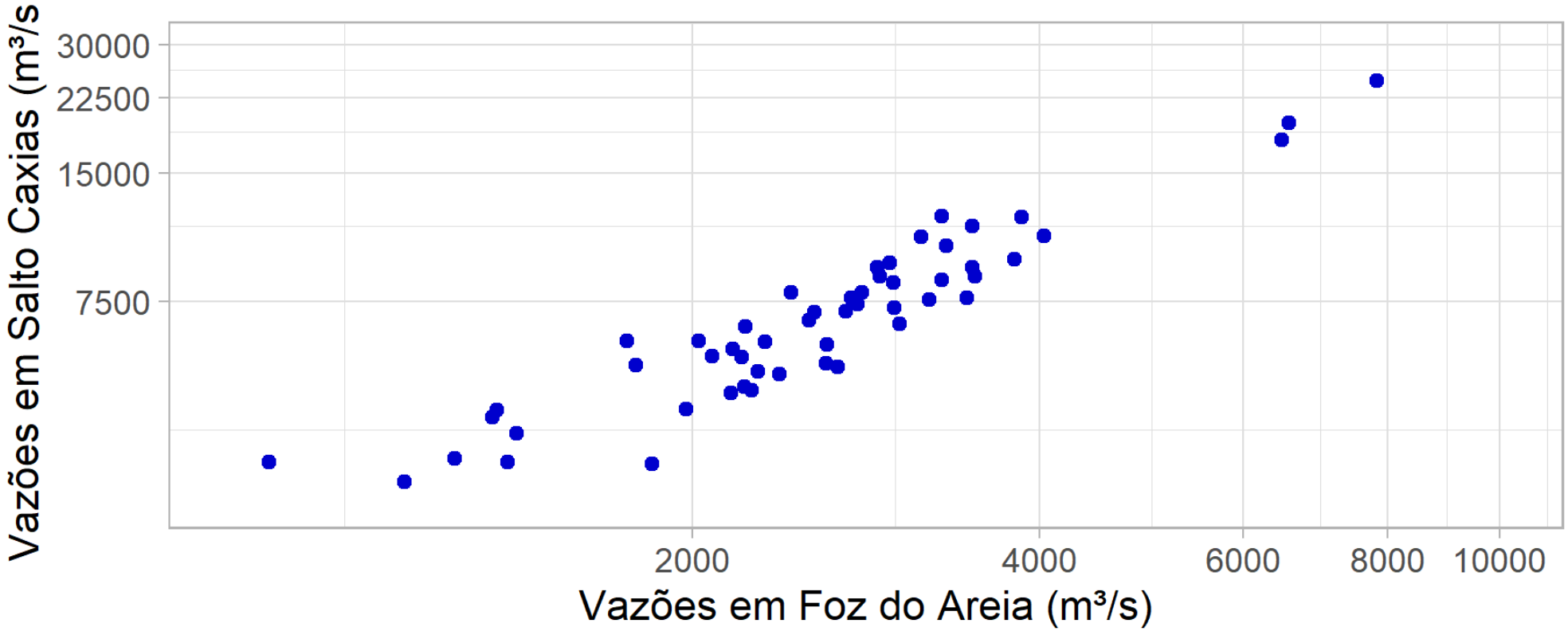
Fonte: Helsel et al. (2020, p. 428)

Análise preliminar de dados | escalas

Escalas transformadas

valores dos eixos devem manter a escala original dos dados

ex.: transformação logarítmica



APRESENTAÇÃO DOS TRABALHOS

Análise Preliminar de Dados

estatísticas descritivas

Análise preliminar de dados | estatísticas descritivas

Estatísticas descritivas: métodos para resumir um conjunto de dados

Provêm informações importantes para inferir aspectos da população

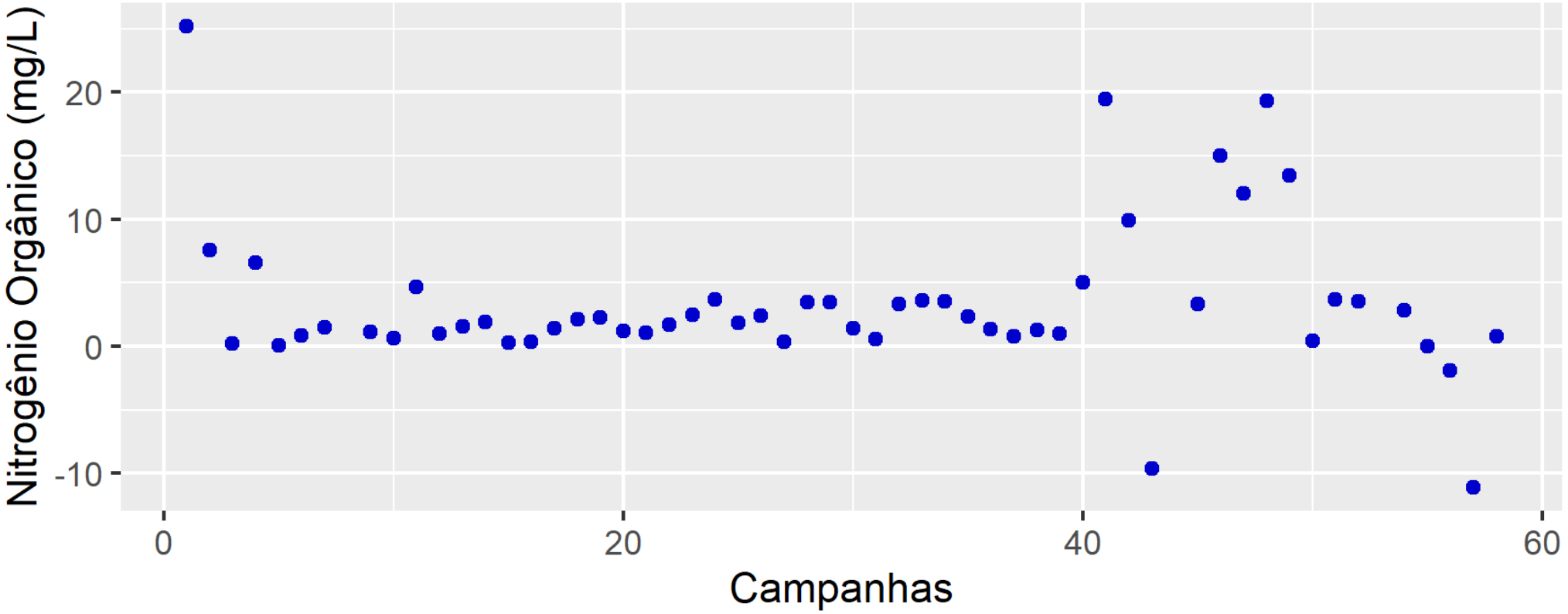
- tendência central
- variabilidade
- assimetria

Prioridade para medidas robustas

- robustez estatística: técnica que funciona bem para amostras com diferentes características

Análise preliminar de dados | estatísticas descritivas

Para os próximos exemplos será utilizada a séries de Nitrogênio Orgânico (NOrg) no posto IG3 (rio Iguaçu), compreendida entre jun./05 e ago./17



Análise preliminar de dados | caracterização dos dados

Caracterização dos dados

Tamanho total da série

pode ou não incluir falhas na observação

falhas tipicamente representadas por códigos: -999, NA, etc.

Quantidade de falhas

expresso tipicamente em percentual

Tamanho da série sem falhas

série efetivamente usada para o cálculo das estatísticas

Análise preliminar de dados | tendência central

Medidas de tendência central

Média aritmética ($\bar{X}$):

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

onde

X_i observação no instante i

n tamanho da amostra

Métrica pouco robusta por sofrer com a influência de *outliers*.

Análise preliminar de dados | tendência central

Mediana ($\tilde{X}$): métrica robusta de tendência central (imune a *outliers*)

equivale ao percentil 50 (P_{50})

Ordenar a série de forma crescente: $X(1) < X(2) < \cdots < X(n)$

Depois:

$$\tilde{X} = \begin{cases} X\left(\frac{n+1}{2}\right), \text{ para n ímpar} \\ \frac{1}{2}\left[X\left(\frac{n}{2}\right) + X\left(\frac{n}{2}+1\right)\right], \text{ para n par} \end{cases}$$

Análise preliminar de dados | variabilidade

Medidas de variabilidade

Variância (s^2):

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

onde

X_i	observação no instante i
$\bar{X}$	média da amostra
n	tamanho da amostra

Métrica pouco robusta por sofrer com a influência de *outliers*.

Análise preliminar de dados | variabilidade

Desvio padrão (s):

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

onde

s^2 variância da amostra

Preferível quando se quer expressar a estatística na unidade original da variável

Métrica pouco robusta por sofrer com a influência de *outliers*.

Análise preliminar de dados | variabilidade

Amplitude interquartil (IQR): métrica robusta de variabilidade

- mede a variabilidade de 50% dos dados centrais da amostra
- elimina a influência de 25% dos dados em cada extremidade

$$IQR = P_{75\%} - P_{25\%}$$

onde

P_j percentil referente aos valores menores ou iguais a j na amostra ordenada

$$P_j = X_{(n+1)\cdot j}$$

Para $P_{75\%}$ e $P_{25\%}$, $j = 0{,}75$ e $0{,}25$, respectivamente

Análise preliminar de dados | variabilidade

Amplitude interquartil (IQR): (cont.)

Ex.: seja a amostra {11, 9, 4, 2, 8, 11, 12}. Ordenando-a, tem-se:

2	4	8	9	11	11	12
1º	2º	3º	4º	5º	6º	7º

$$P_{25\%}: X_{(7+1)\cdot 0{,}25} = X_2 \rightarrow 4$$
$$P_{75\%}: X_{(7+1)\cdot 0{,}75} = X_6 \rightarrow 11$$
$$\therefore IQR = 11 - 4 = 7$$

Quando os índices não resultam em valores inteiros, é preciso interpolar

Análise preliminar de dados | variabilidade

Desvio absoluto da mediana (MAD):

$$MAD(X) = \left|\tilde{d}_i\right|$$

onde

$\tilde{d}_i$ mediana de d_i

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

onde

X_i observação no instante i

$\tilde{X}$ mediana da amostra

Análise preliminar de dados | variabilidade

Coeficiente de variação (CV): medida adimensional de dispersão

útil para comparar o nível de dispersão de séries com diferentes escalas

$$CV = \frac{s}{\bar{X}} (\times 100)$$

onde

$\bar{X}$ média da amostra

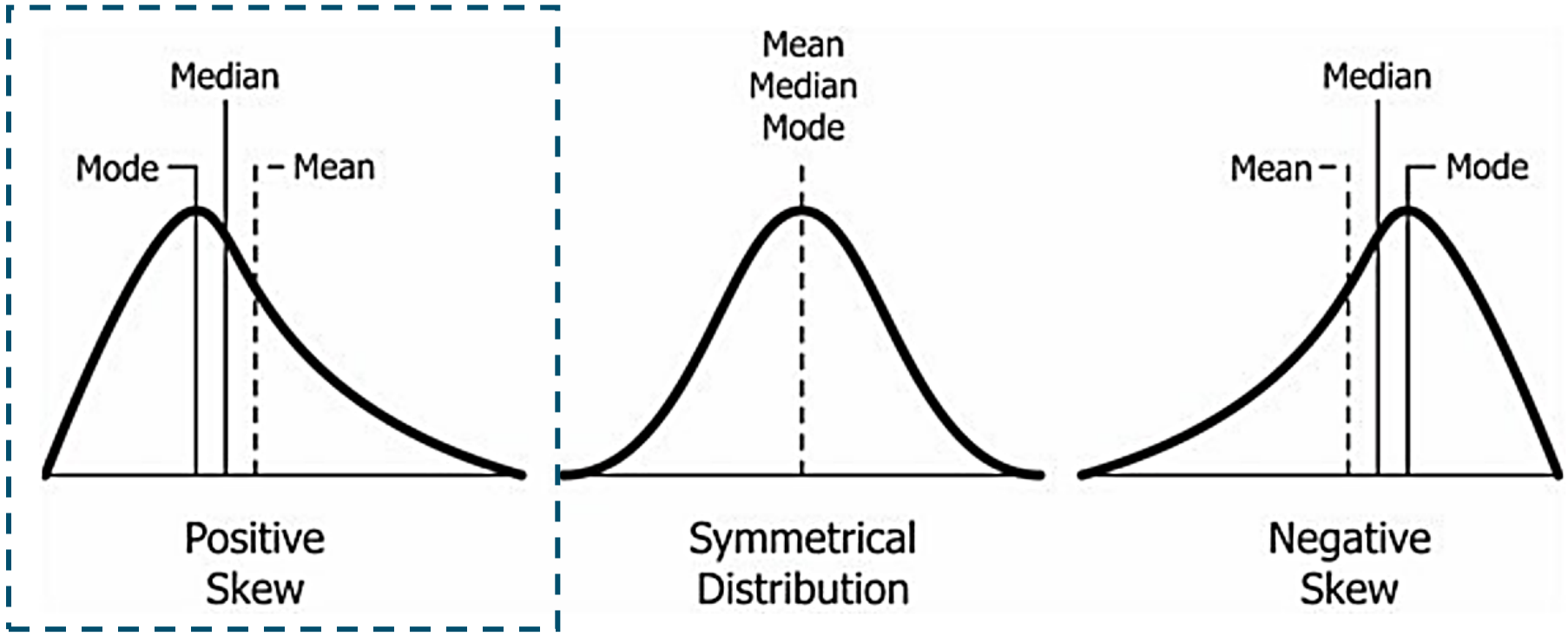
s desvio padrão da amostra

$CV \leq 15\%$	baixa dispersão
$15\% < CV \leq 30\%$	média dispersão
$CV > 30\%$	alta dispersão

Análise preliminar de dados | assimetria

Medidas de assimetria

A assimetria positiva é característica inerente de variáveis naturais



Fonte: https://upload.wikimedia.org/wikipedia/commons/c/cc/Relationship_between_mean_and_median_under_different_skewness.png

Análise preliminar de dados | assimetria

Em dados com assimetria positiva, tanto a média quanto o desvio padrão (ou variância) são métricas pouco precisas

- média é sempre maior que a mediana
- desvio padrão (ou variância) são tendenciosos

Para esses casos, métricas adicionais (robustas) precisam ser incluídas na análise exploratória dos dados

Assimetria também indica a não normalidade dos dados

- cuidados no uso de técnicas estatísticas paramétricas
- (mais detalhes no decorrer do curso)

Análise preliminar de dados | assimetria

Coeficiente de assimetria (g): medida clássica de assimetria

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^3}{s^3}$$

onde

X_i	observação no instante i
$\bar{X}$	média da amostra
s	desvio padrão da amostra
n	tamanho da amostra

Métrica pouco robusta por sofrer com a influência de *outliers*.

Análise preliminar de dados | assimetria

Em amostras com tamanhos inferiores a 100 elementos, g pode ser altamente tendencioso

> tendenciosidade estatística: diferença entre o valor amostral e o populacional de uma estimativa

A assimetria amostral tende a ser subestimada em relação à assimetria populacional

Somente quando a assimetria populacional é nula, g é uma estimativa não tendenciosa

Análise preliminar de dados | assimetria

Assimetria quartílica (*quartile skew* – qs): medida robusta de assimetria

$$qs = \frac{(P_{75\%} - P_{50\%}) - (P_{50\%} - P_{25\%})}{(P_{75\%} - P_{25\%})}$$

onde

P_j percentil referente aos valores menores ou iguais a j na amostra ordenada

$$P_j = X_{(n+1)\cdot j}$$

Para $P_{75\%}$, $P_{50\%}$ e $P_{25\%}$, $j = 0{,}75$, $0{,}50$ e $0{,}25$, respectivamente

Análise preliminar de dados | aplicação

Aplicação em R: estatísticas descritivas

Caracterização	Tamanho da série	58 elementos
	Falhas	5,2%
Tendência central	Média	3,39 mg/L
	Mediana	1,88 mg/L
Variabilidade	Variância	36,11 $(mg/L)^2$
	Desvio Padrão	6,01 mg/L
	Amplitude interquartil	2,83 mg/L
	Desvio absoluto da mediana	2,16 mg/L
	Coef. de variação	56%
Assimetria	Coef. de assimetria	1,41
	Assimetria interquatílica	0,22

Revisão

Gráficos são a forma mais eficiente de comunicação e divulgação

- priorizar o que se quer mostrar
- cuidar com cores, eixos e escalas

Estatísticas descritivas são úteis para melhor entendimento da amostra

- medidas de tendência central, variação e assimetria
- assimetria inerente às variáveis ambientais pode distorcer estatísticas clássicas
- priorizar medidas robustas

Estatística Aplicada a Ciências Ambientais

Daniel Detzel
detzel@ufpr.br