Estatística Aplicada a Ciências Ambientais

Análise Preliminar de Dados (pt.1)

Daniel Detzel
detzel@ufpr.br

Introdução | generalidades

Apresentações individuais (de quem faltou...)

- nome
- área de concentração
- conhecimento de estatística
- conhecimento de programação (R, Python, Matlab, etc.)
- o que espera da disciplina

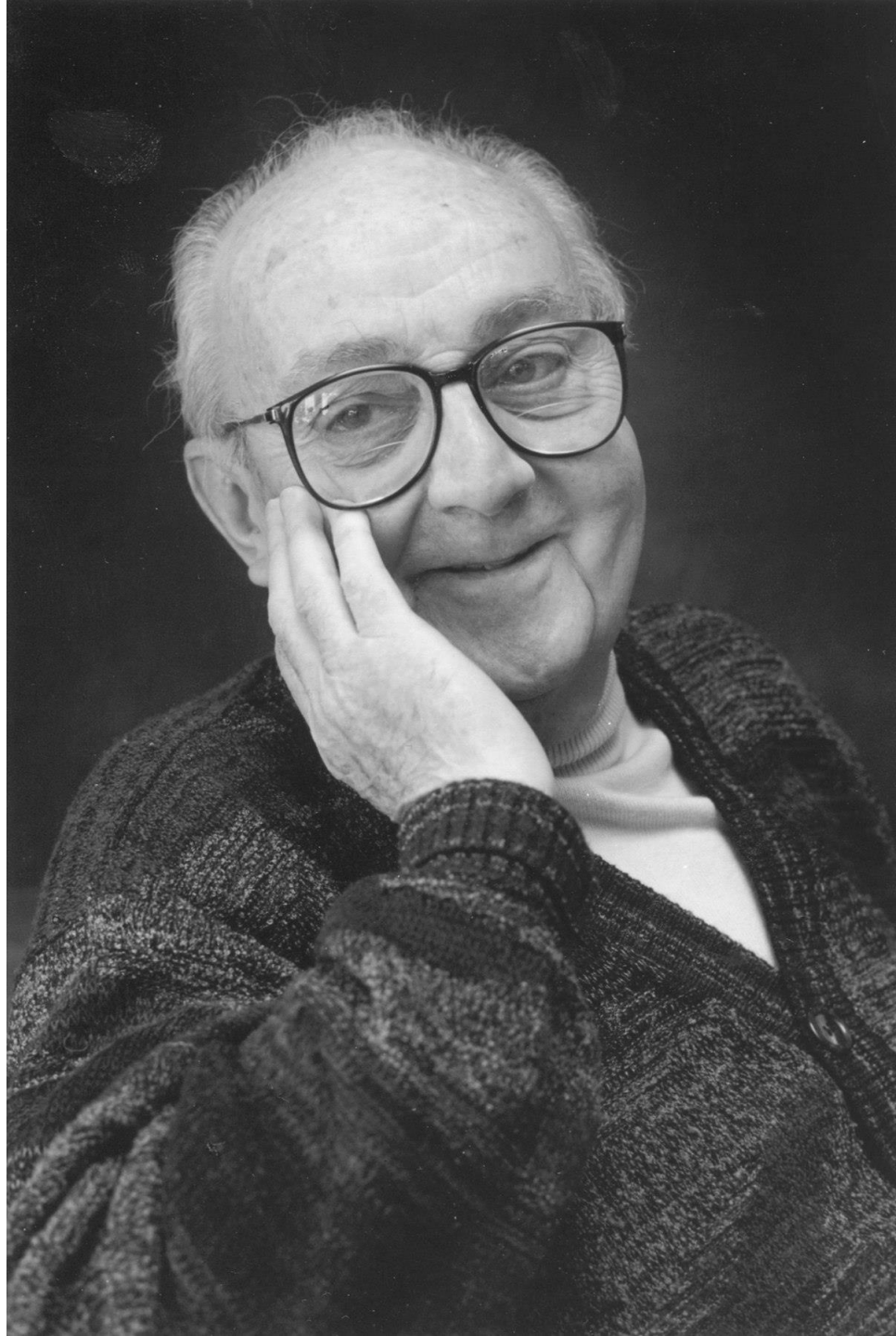
Análise Preliminar de Dados

características dos dados hidrológicos e de qualidade da água

”

Apaixone-se pelos seus dados, nunca pelos seus modelos.

frase atribuída a George Box



Análise preliminar de dados | características

O que esperar de dados hidrológicos e de qualidade da água?

1. São frequentemente limitados em zero

    dados negativos são raros
    exceções: temperatura do ar, cotas (dependendo do referencial), etc.

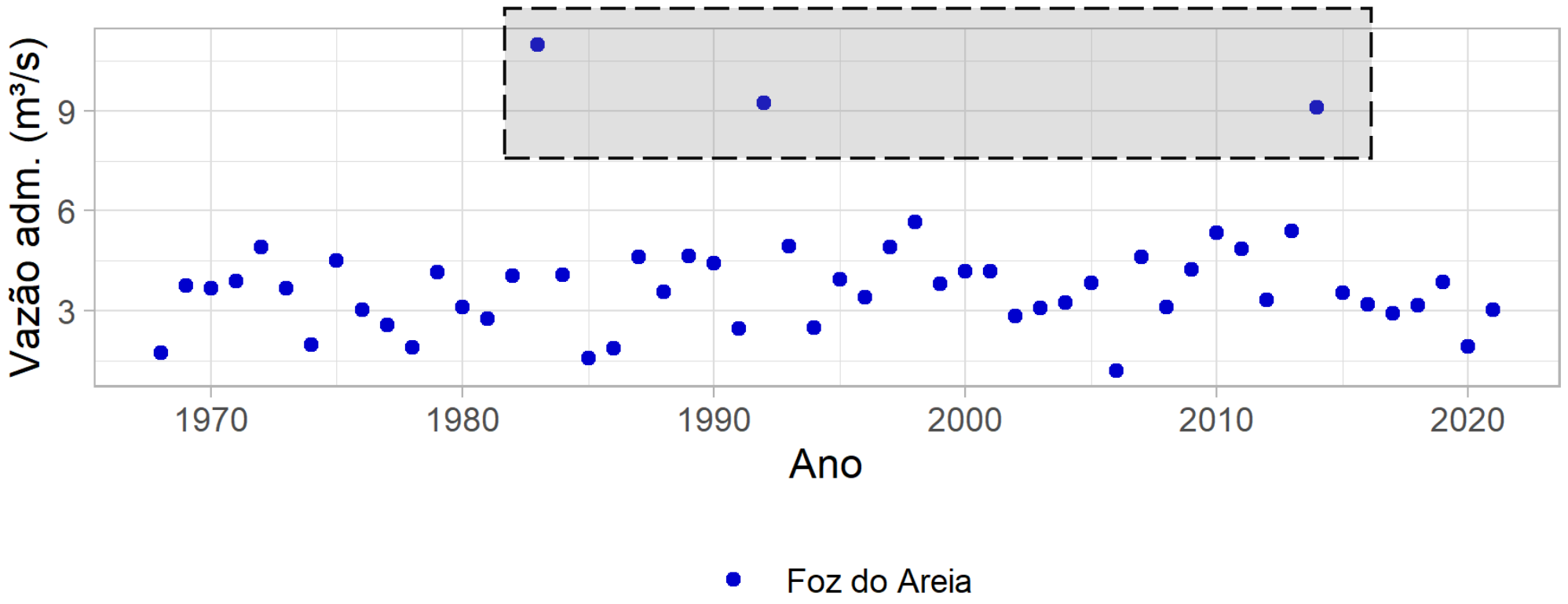
2. Podem apresentar *outliers* (“pontos fora da curva”)

    valores **consideravelmente** maiores ou menores do que o restante da amostra
    resultados de erros nas observações ou ocorrência de eventos extremos
    recomenda-se discernimento na decisão se o valor é ou não um *outlier*
    em casos positivos, são removidos da amostra

Análise preliminar de dados | características

2. Podem apresentar *outliers* (“pontos fora da curva”) (cont.)

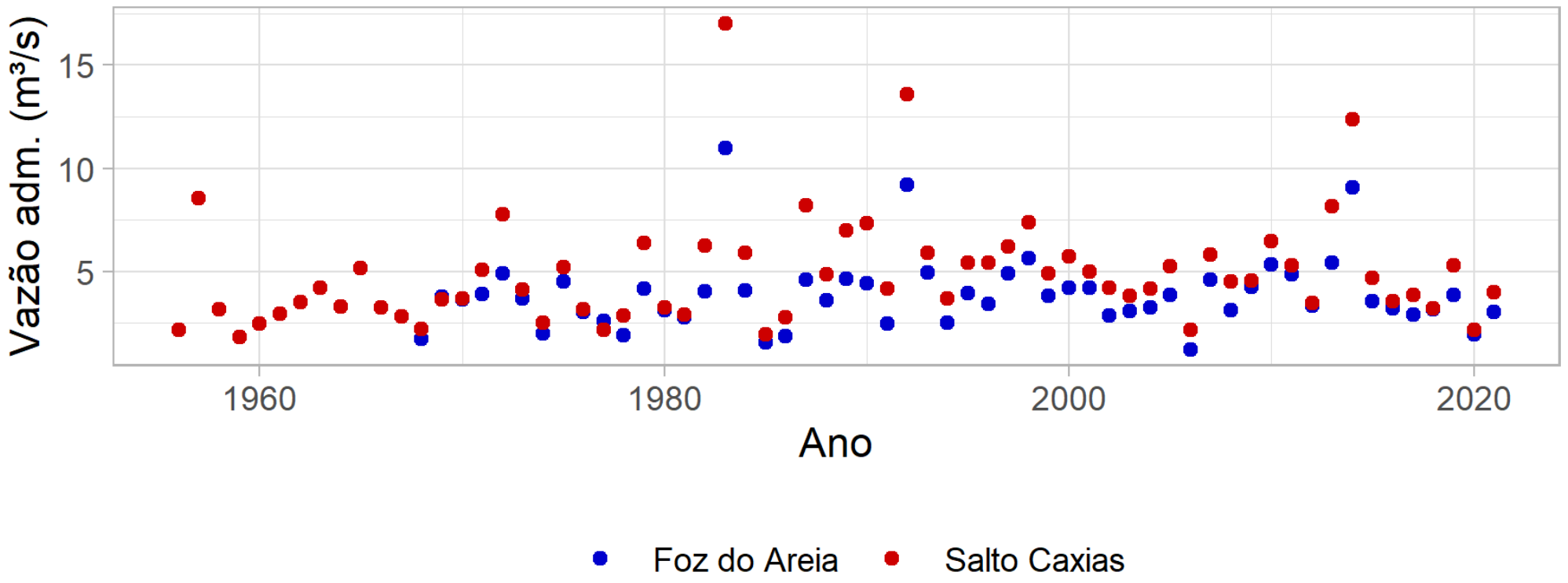
Ex. 1: Vazões máximas anuais em Foz do Areia (rio Iguaçu)



Análise preliminar de dados | características

2. Podem apresentar *outliers* (“pontos fora da curva”) (cont.)

Ex. 1: Vazões máximas anuais em Foz do Areia e Salto Caxias (ambas no rio Iguaçu)

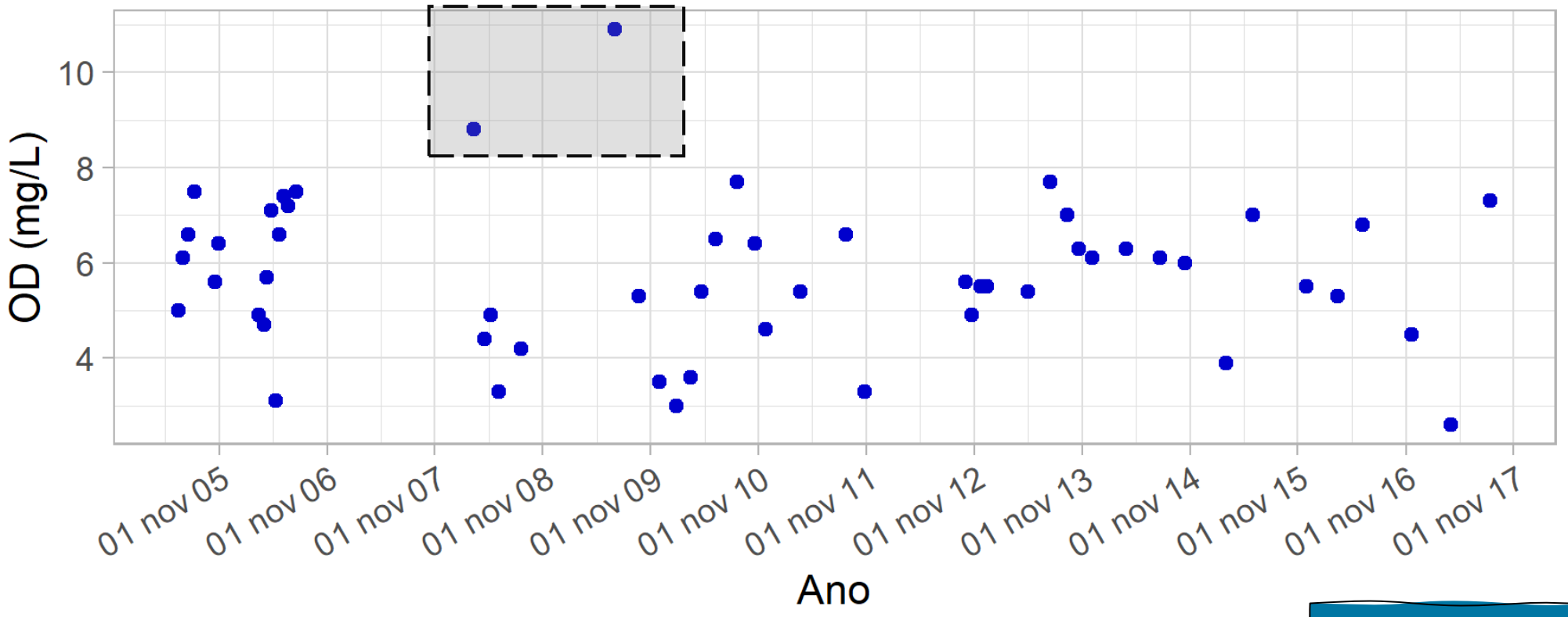


Não é *outlier*!

Análise preliminar de dados | características

2. Podem apresentar *outliers* (“pontos fora da curva”) (cont.)

Ex. 2: Oxigênio Dissolvido no posto IG1 (rio Iguaçu)



Possível *outlier*!

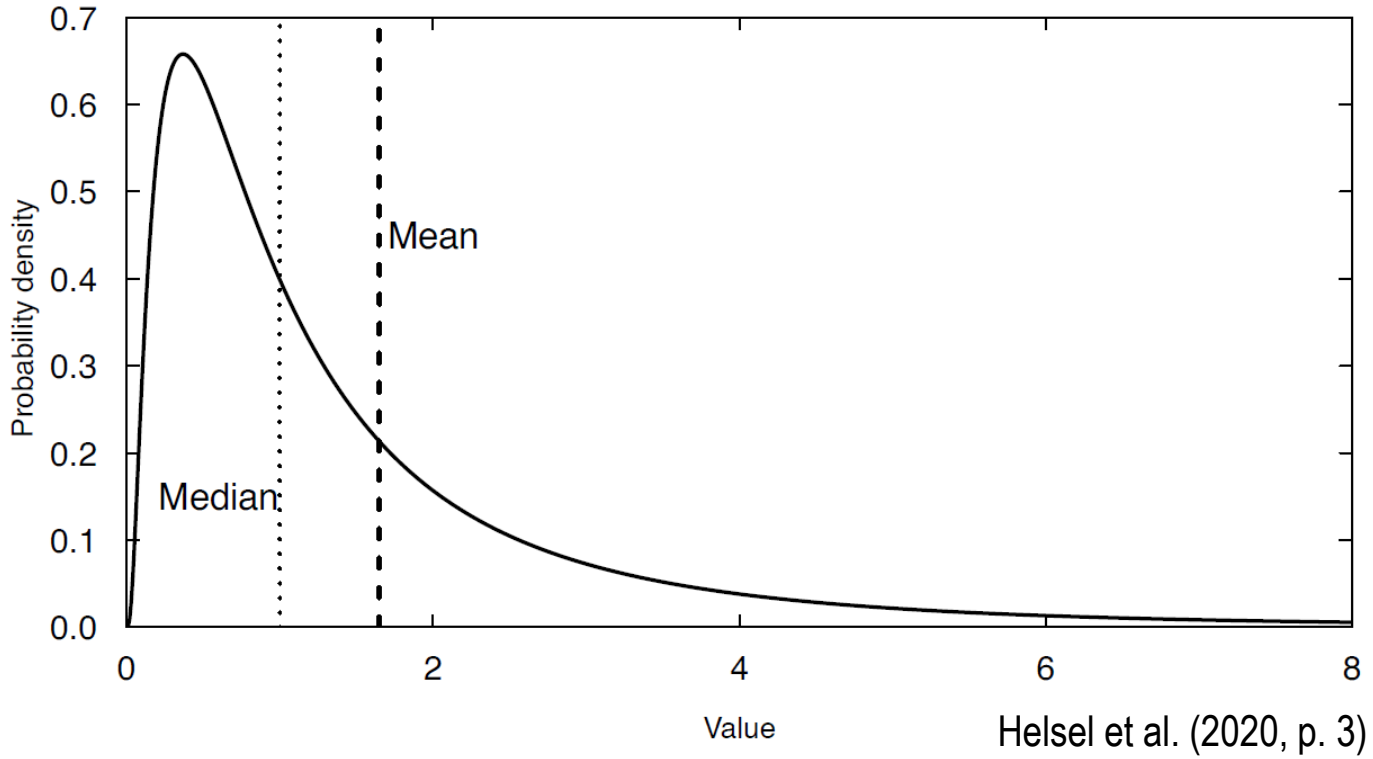
Análise preliminar de dados | características

O que esperar de dados hidrológicos e de qualidade da água? (cont.)

3. Apresentam assimetria positiva

valores afastados do centro da distribuição

acúmulo de valores do lado esquerdo da distribuição

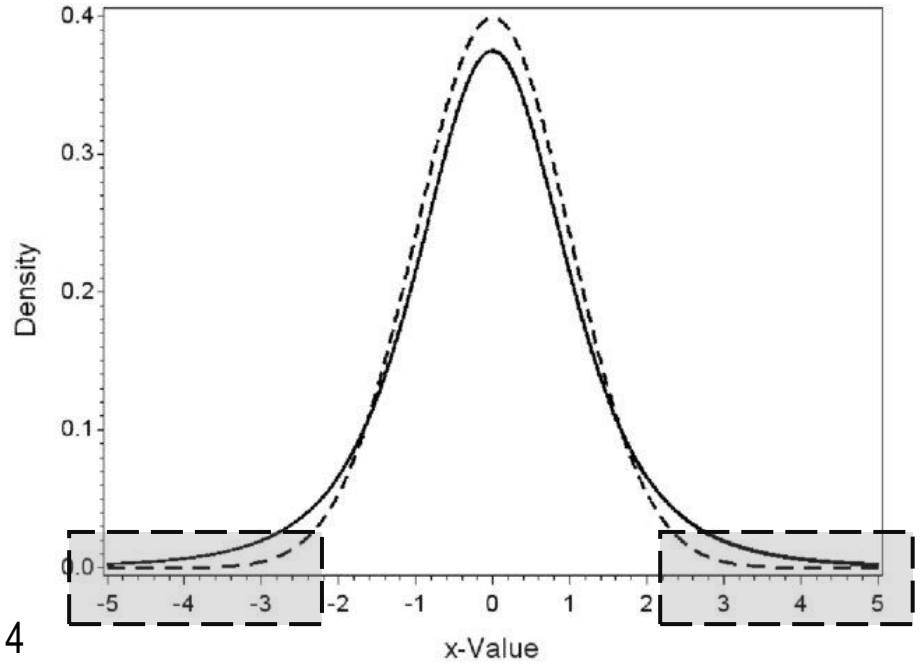


Helsel et al. (2020, p. 3)

Análise preliminar de dados | características

O que esperar de dados hidrológicos e de qualidade da água? (cont.)

4. Apresentam distribuições não normais
    - cuidados na aplicação de técnicas estatísticas que assumem normalidade
    - é possível trabalhar com transformações nos dados
    - mesmo em valores com assimetria nula, cuidados devem ser tomados com as caudas da distribuição (valores extremos)



Fonte: https://doi.org/10.1515/cttr-2017-0014

Análise preliminar de dados | características

O que esperar de dados hidrológicos e de qualidade da água? (cont.)

5. Variáveis limitadas (censuradas)
   - consequência do processo natural, ou da forma de obtenção/medição
   - precisão dos equipamentos utilizados na amostragem

6. Padrões sazonais
   - valores maiores ou menores em determinados períodos do ano
   - os períodos podem variar (meses, estações, semestres, etc.)

7. Dependência de outras variáveis
   - importante para explicar variações e relações causa-efeito
   - ex.: parâmetros de qualidade da água dependem da vazão

Análise preliminar de dados | características

O que esperar de dados hidrológicos e de qualidade da água? (cont.)

8. Persistência (autocorrelação)

dependência de um valor em relação às observações em seu entorno

no tempo: indica correlação serial

no espaço: indica proximidade geográfica

propriedade determinante para a representatividade estatística de uma amostra

cuidados no emprego de técnicas que requerem independência

Dica de leitura (disponível no material suplementar da disciplina):
Mandelbrot, B.B.; Wallis, J.R. Noah, Joseph, and Operational Hydrology. Water Resources Research, v. 4, n. 3, 1968.

Análise preliminar de dados | características

O que esperar de dados hidrológicos e de qualidade da água? (cont.)

9. Erros nas medições

sempre presentes, não importa o cuidado na obtenção das amostras

erros aleatórios podem ser estimados e reduzidos

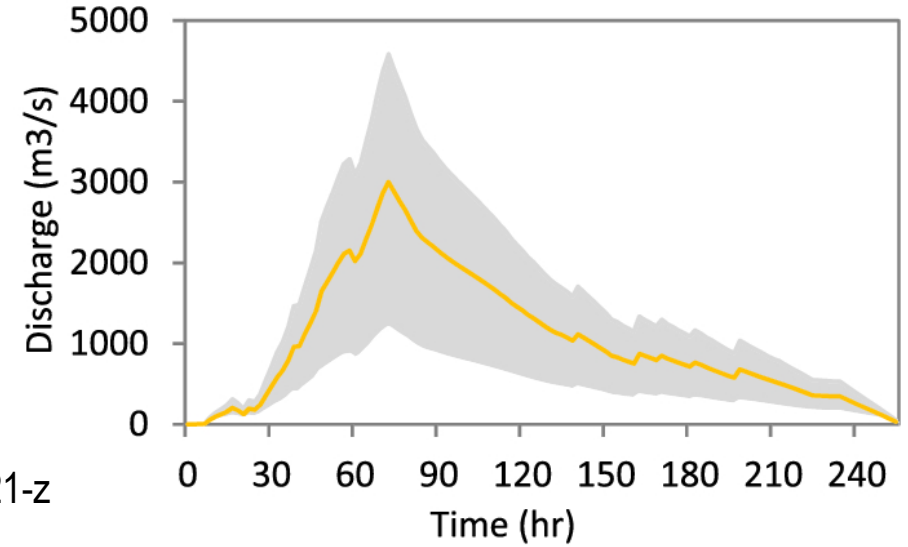
ex.: repetição de experimentos (duplicatas)

erros sistemáticos requerem mais cuidados

ex.: sondas sem calibração; alteração nas condições de amostragem

contribuem com a incerteza das estimativas

proporcionais à magnitude da variável



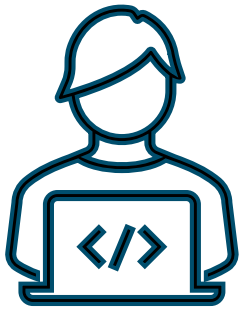
Fonte: https://doi.org/10.1007/s11069-021-04621-z

Análise Preliminar de Dados

representação gráfica de dados

Análise preliminar de dados | representação gráfica

Configuração RStudio



1. Instalar pacote “ggplot2”

```
install.packages("ggplot2")
```

2. Fazer a leitura da(s) série(s) escolhida(s)

```
nome <- read.csv('nomeDoArquivo.csv')
```

para arquivos em xlsx, sugestão de pacote “readxl”

```
install.packages("readxl")
nome <- read_excel('nomeDoArquivo.xlsx')
```

Análise preliminar de dados | representação gráfica

Gráficos são eficientes em resumir informações das séries em estudo

Representações gráficas fazem parte da Análise Exploratória de Dados, que se refere ao primeiro olhar sobre as variáveis

- permitem identificar as características mencionadas anteriormente
- direcionam as análises seguintes

São também úteis para comunicar as informações via relatórios, artigos ou apresentações

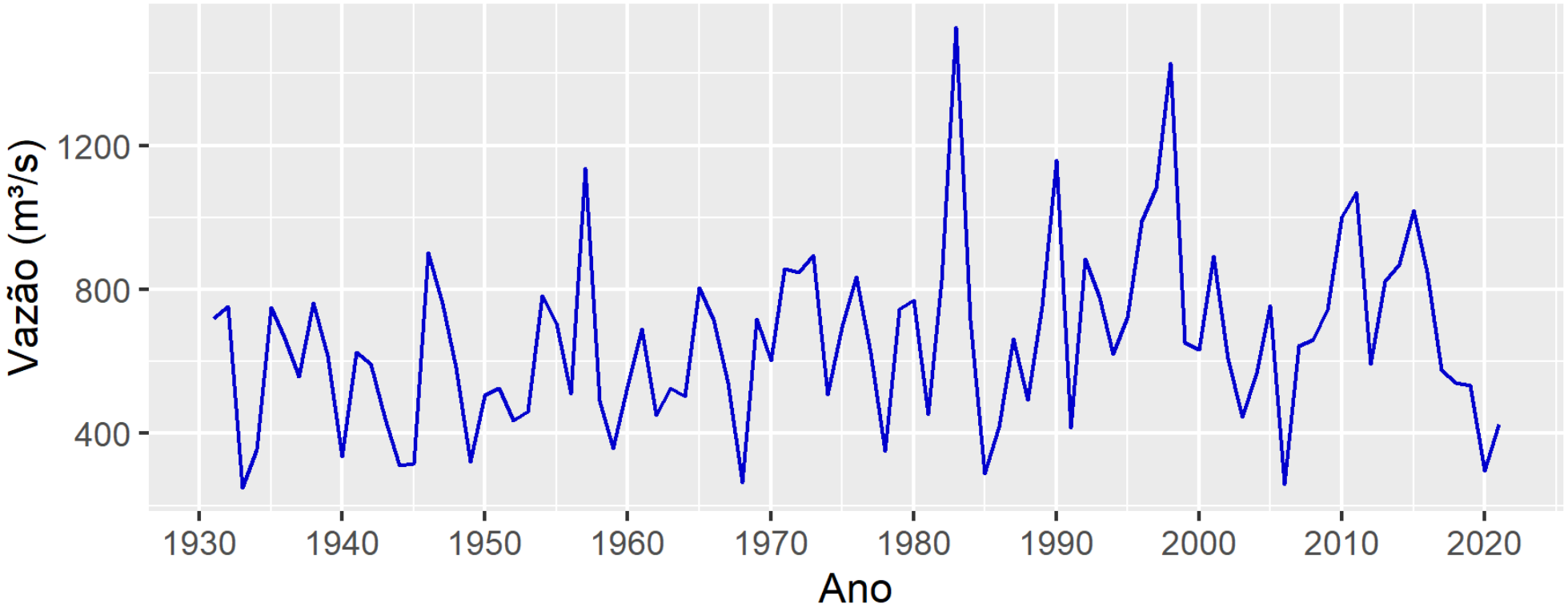
Para os próximos exemplos será utilizada a séries de vazões médias anuais em Foz do Areia (rio Iguaçu), compreendida entre 1931 e 2021

Análise preliminar de dados | série temporal

Série temporal

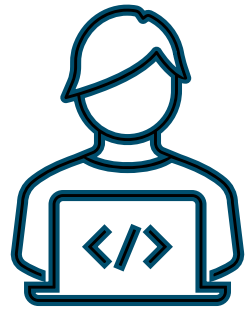
Simples plotagem dos dados

útil para avaliar as caraterísticas gerais da amostra



Análise preliminar de dados | série temporal

No R, usando ggplot2



A variável deve ser um `data frame`. Para montar um:

```
nome <- data.frame(anos = dataSerie,
                   valores = valorSerie)
```

Para plotar:

```
ggplot(nome, aes(x=anos, y=valores)) +
     geom_line(colour="blue3”) +
     labs(x="Ano",y=“Variável (unidade)") +
     theme_gray()
```

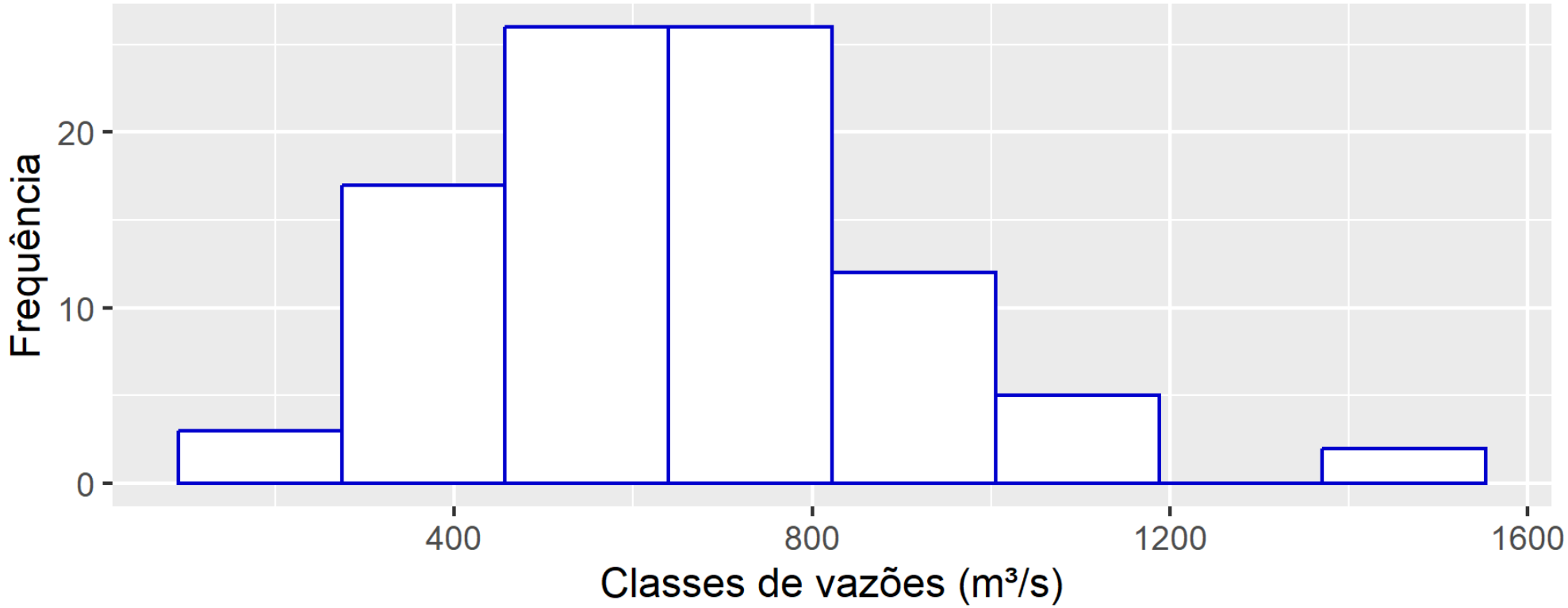
Obs.: nos códigos, variáveis em pretos são necessários para as funções. Valores em azul são escolhas do usuário.

Análise preliminar de dados | histograma

Histograma

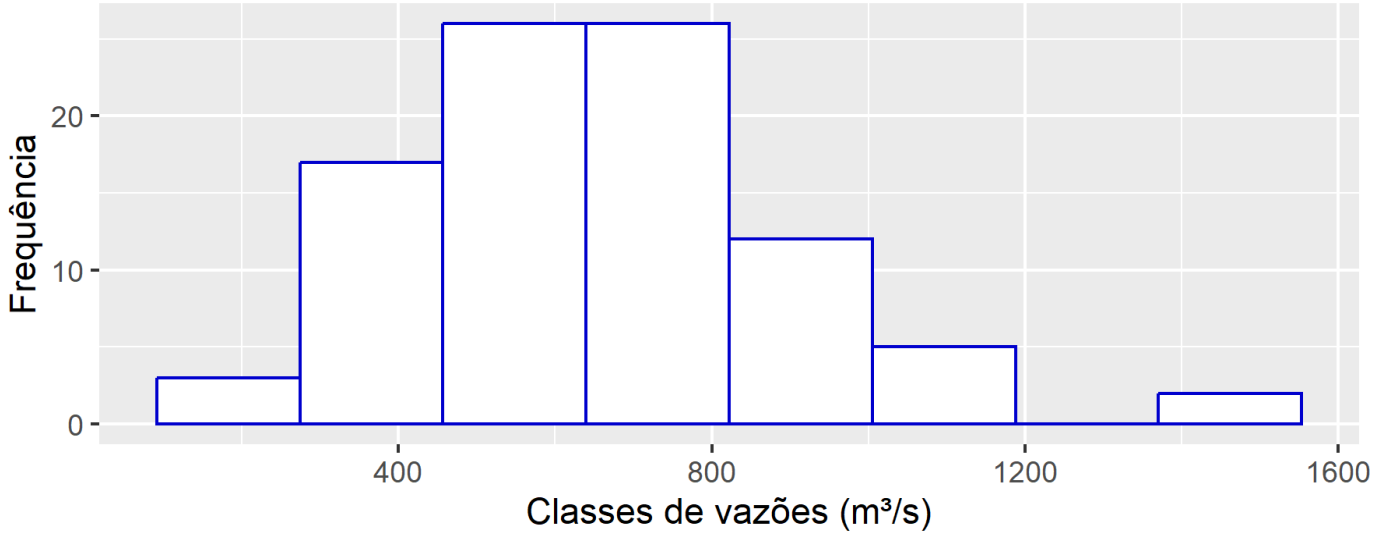
Representa a distribuição dos dados de acordo com classes

útil para avaliar a tendência central, variabilidade e assimetria

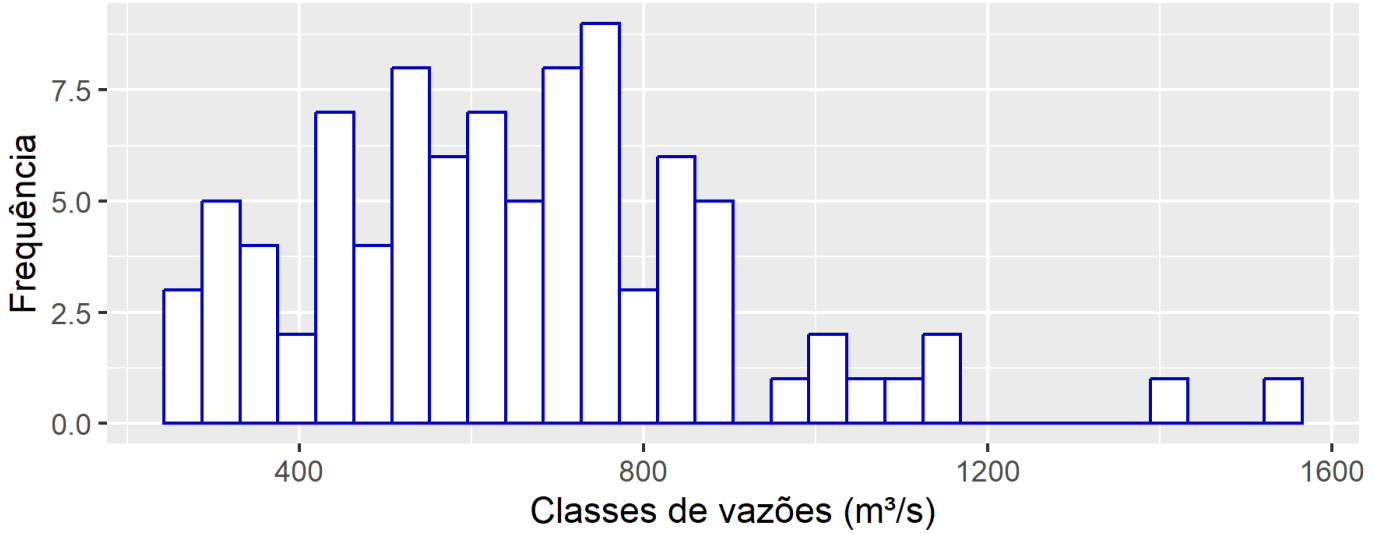


Análise preliminar de dados | histograma

Cuidado: a aparência depende do número de classes



8 classes



30 classes

Análise preliminar de dados | histograma

Critérios para definição do número k de classes em uma amostra de n elementos

1. Iman e Conover: menor inteiro que satisfaça $2^k \geq n$

   para o exemplo, $n = 91$

   se $k = 6 \rightarrow 2^6 = 64$

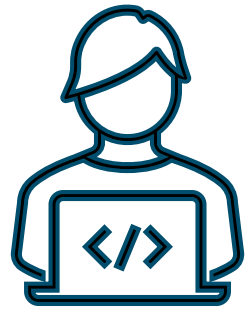
   se $k = 7 \rightarrow 2^7 = 128 \quad \therefore k = 7$

2. Sturges: $k = 1 + 3{,}3 \log n$

   para o exemplo, $n = 91 \quad \therefore k \cong 8$

Análise preliminar de dados | histograma

No R, usando ggplot2



A variável deve ser um `data frame`

```
ggplot(nome, aes(x=var)) +
     geom_histogram(bins=k)
```

Parâmetros opcionais para edição

```
ggplot(nome, aes(x=var)) +
     geom_histogram(colour="blue3", fill="white", bins=k) +
     labs(x="Classes da variável (unidade)",y="Frequência") +
     theme_gray()
```

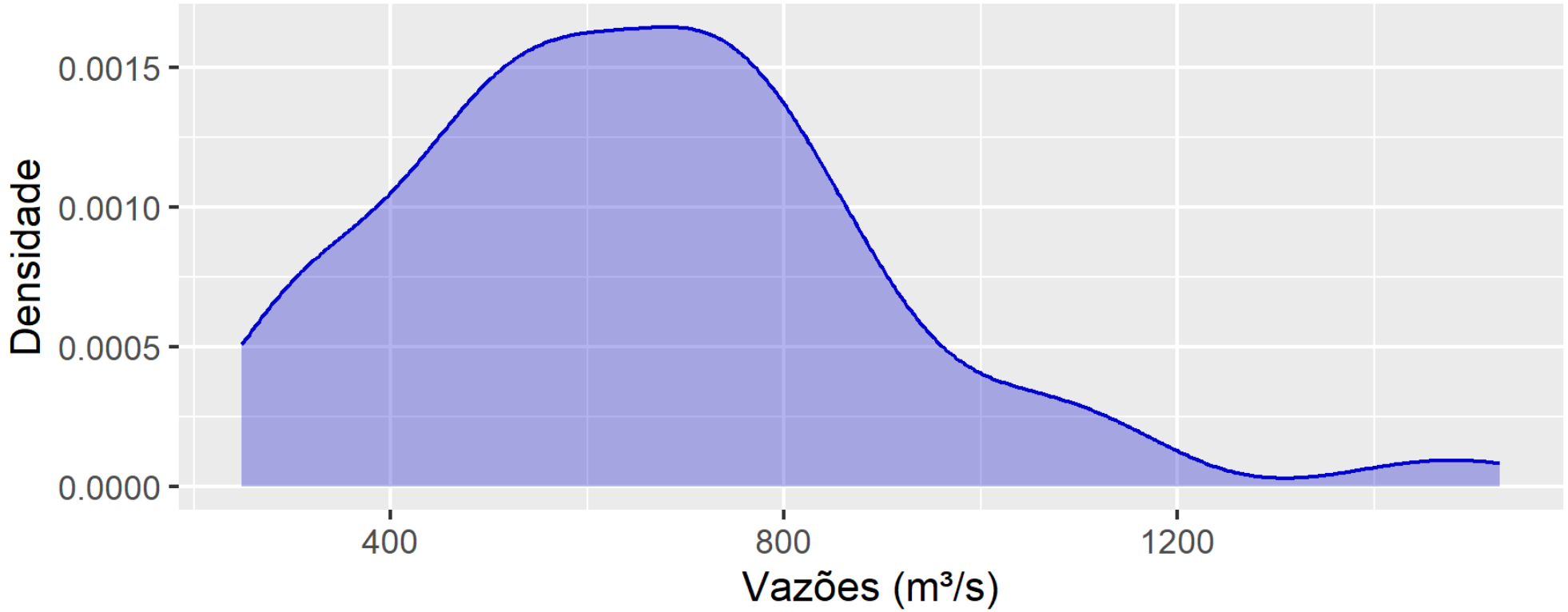
Análise preliminar de dados | densidade

Densidade

Versão suavizada do histograma

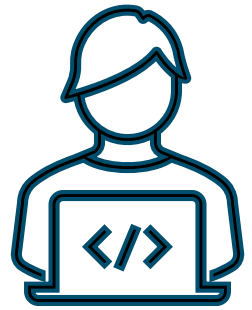
útil para avaliar a tendência central, variabilidade e assimetria

densidade: probabilidade para um intervalo de valores do eixo x



Análise preliminar de dados | densidade

No R, usando ggplot2



A variável deve ser um `data frame`

```
ggplot(nome, aes(x=var)) +
     geom_density()
```

Parâmetros opcionais para edição

```
ggplot(nome, aes(x=var)) +
     geom_density(colour="blue3", fill="white", alpha=0.3) +
     labs(x="Variável (unidade)",y="Densidade") +
     theme_gray()
```

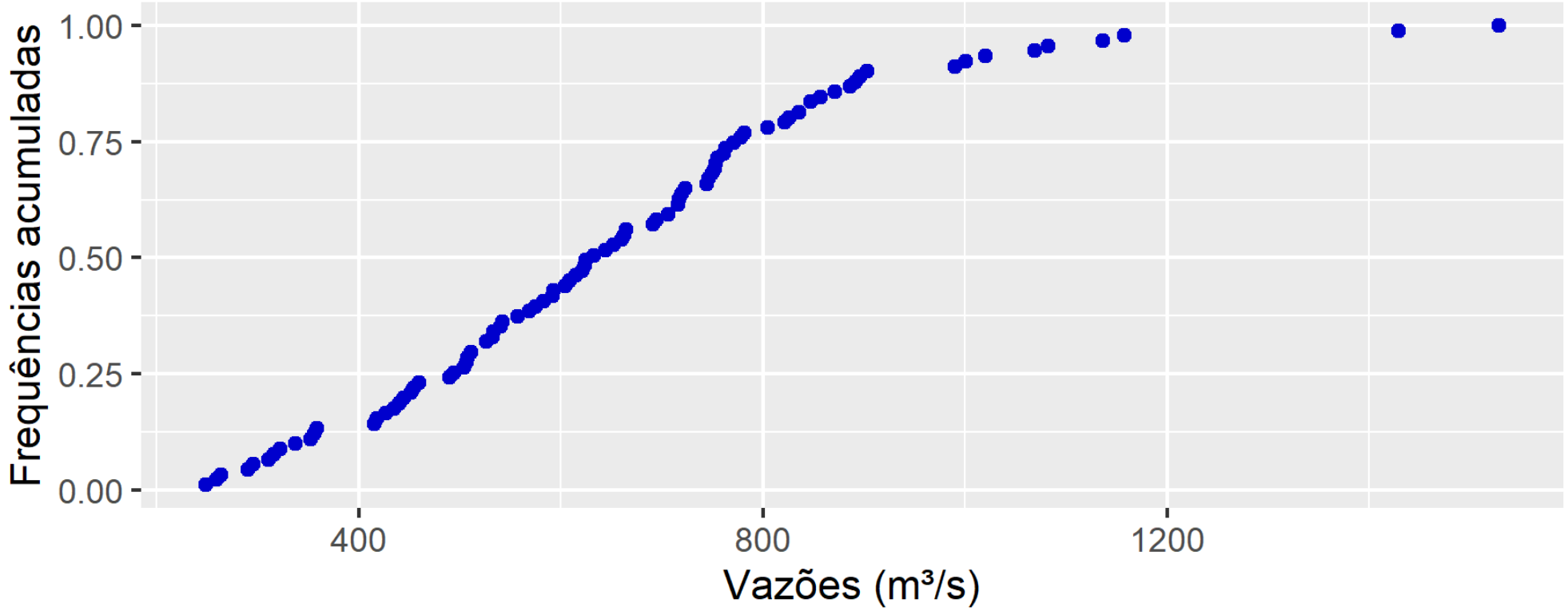
Análise preliminar de dados | frequência acumulada

Frequência acumulada

Expressa a probabilidade de excedência dos valores da amostra

contém todas as n observações

mais preciso do que histograma/densidade para avaliar observações individuais



Análise preliminar de dados | frequência acumulada

Obtida a partir dos seguintes passos:

1. Ordenar a amostra da menor para a maior
2. Associar ranques i em valores crescentes:
   menor valor: $i = 1$
   maior valor: $i = n$
3. Obter as posições de plotagem, usando um método apropriado. Ex.:

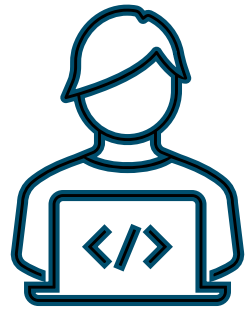
$$f_i = \frac{i}{(n+1)}$$

4. Plotar os dados ordenados (eixo x) com as posições f_i (eixo y)

Obs.: método análogo à curva de permanência!

Análise preliminar de dados | frequência acumulada

No R, usando ggplot2



A variável deve ser um `data frame`

```
ggplot(nome, aes(x=var)) +
     stat_ecdf()
```

Parâmetros opcionais para edição

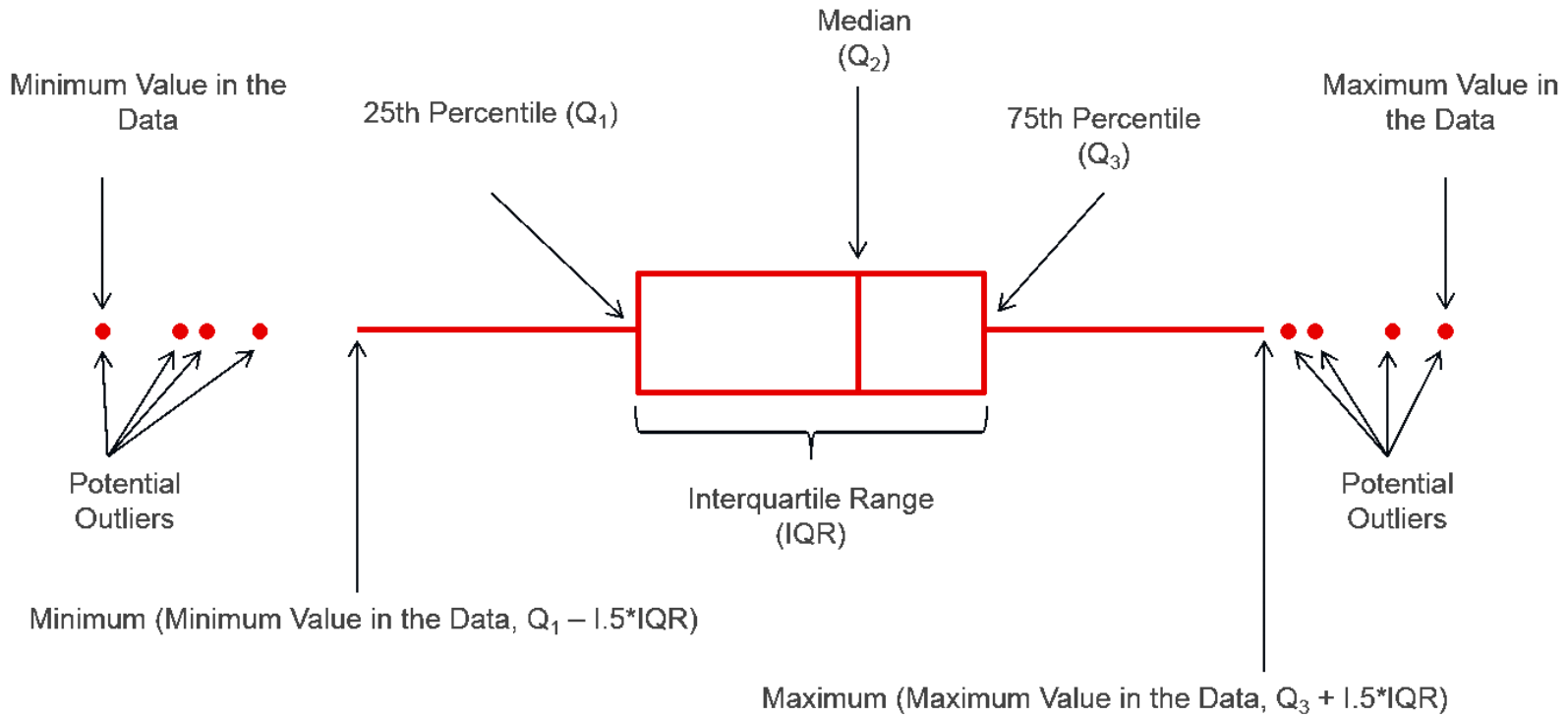
```
ggplot(nome, aes(x=var)) +
     stat_ecdf(geom="point", pad=FALSE, colour="blue3") +
     labs(x="Variável (unidade)",y="Frequências") +
     theme_gray()
```

Análise preliminar de dados | boxplot

Boxplot

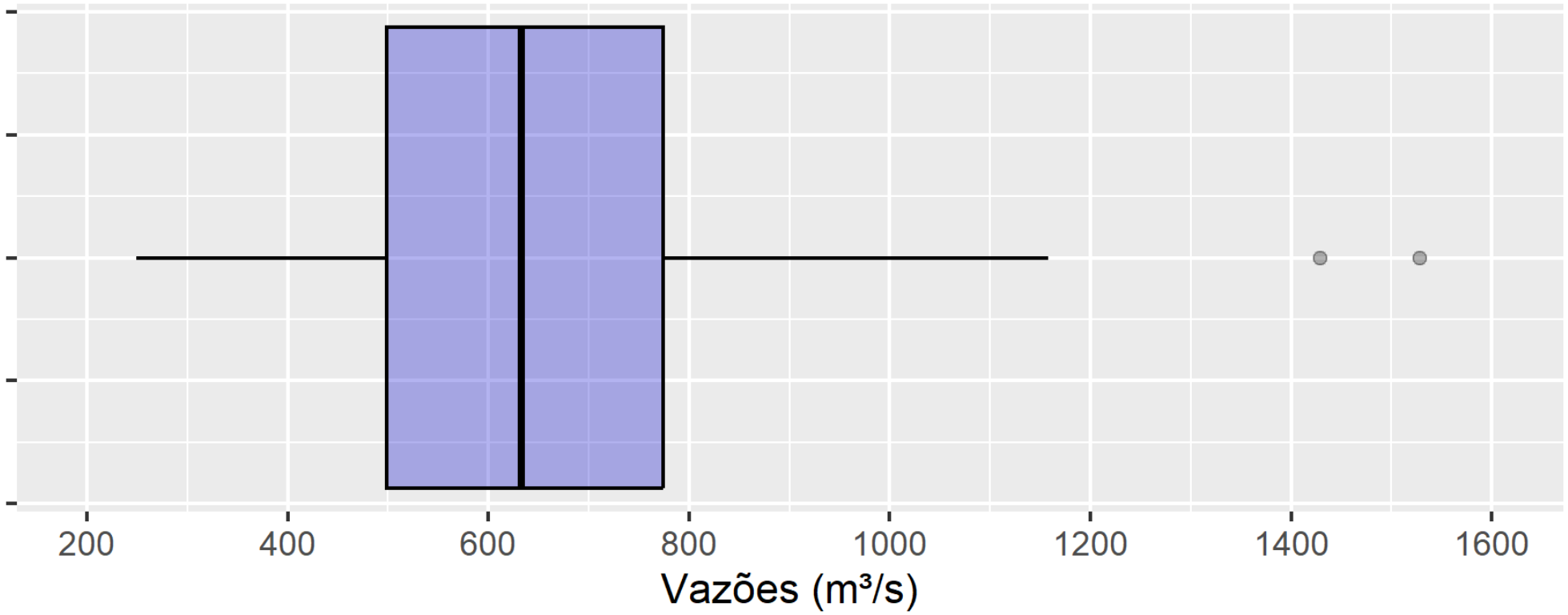
Representação concisa da distribuição de uma amostra

Considerada a maneira mais eficiente de resumir as características de um conjunto de dados



Fonte: https://www.leansigmacorporation.com/box-plot-with-minitab/

Análise preliminar de dados | boxplot



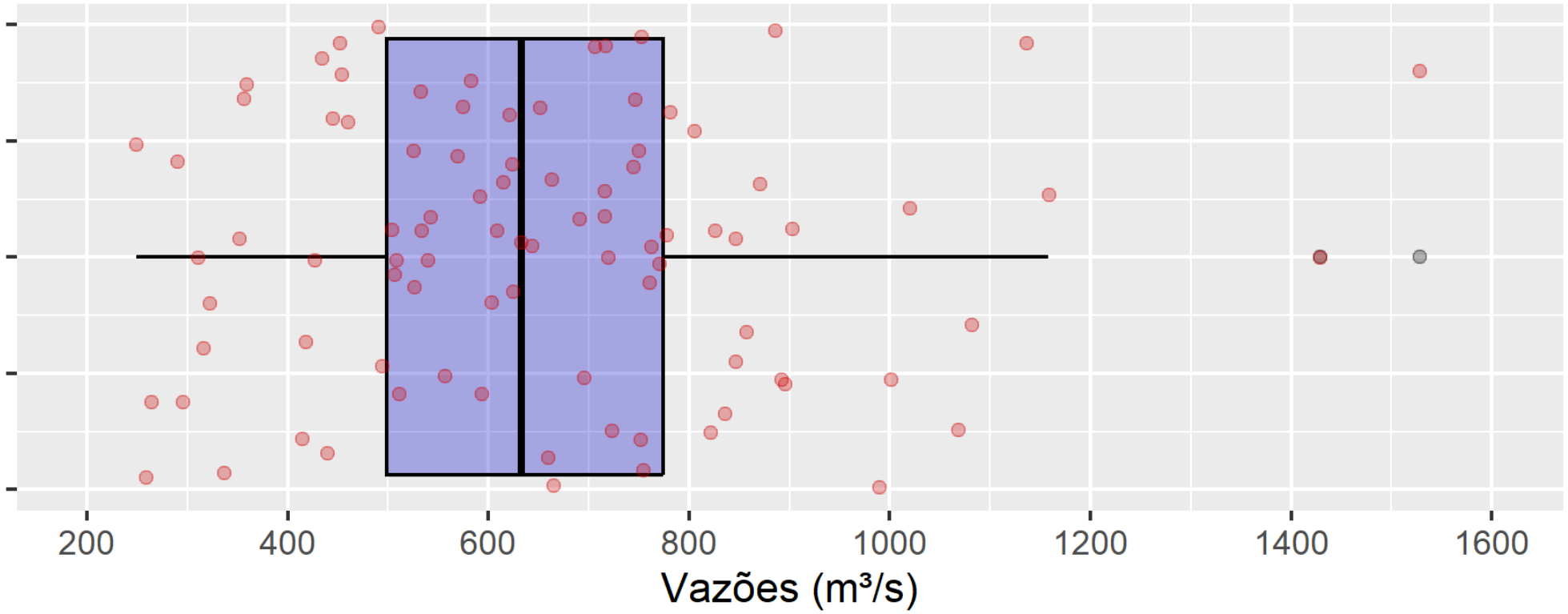
O boxplot tem a limitação de não apresentar a distribuição dos dados

- pode mascarar elementos importantes

Análise preliminar de dados | boxplot

Para amostras não tão grandes: adicionar *jitter* (“perturbações”)

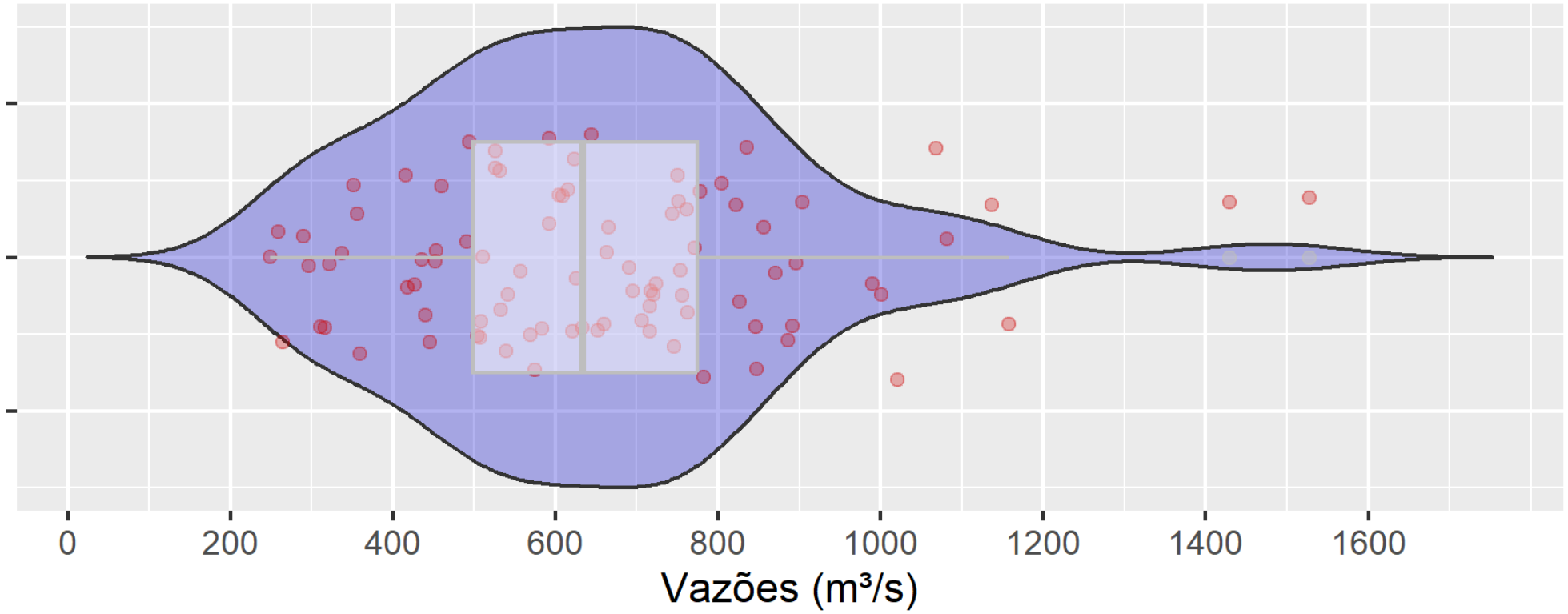
maneira de representar todos os elementos da amostra junto com o boxplot



Análise preliminar de dados | boxplot

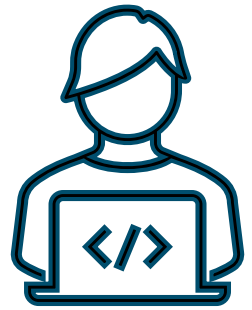
Para amostras grandes: mudar para *violin plot* (“gráfico violino”)

maneira de representar a densidade dos elementos da amostra



Análise preliminar de dados | boxplot

No R, usando ggplot2



A variável deve ser um `data frame`

```
ggplot(nome, aes(x=var,y=var)) +
     geom_boxplot()

ggplot(nome, aes(x=var,y=var)) +
     geom_boxplot() +
     geom_jitter()

ggplot(nome, aes(x=var,y=var)) +
     geom_violin() +
     geom_boxplot() +
     geom_jitter() (opcional)
```

Análise preliminar de dados | representação gráfica

Existem muitas outras opções para elaboração de gráficos

Dicas de consulta:

The R Graph Gallery: https://r-graph-gallery.com/
- um guia com múltiplos tipos de gráficos, suas aplicações e códigos
- prioriza o uso do pacote ‘ggplot2’

ggplot2 *cheat sheets*
- “códigos de trapaça” do pacote ggplot2
- disponíveis online e no material suplementar da disciplina

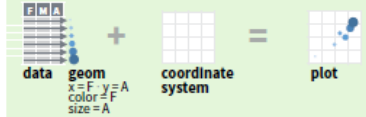
Análise preliminar de dados | representação gráfica

Data Visualization with ggplot2 : : CHEAT SHEET

Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data** set, a **coordinate system**, and geoms—visual marks that represent data points.

To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.



Complete the template below to build a graph.

```
ggplot (data = <DATA>) +                                   required
<GEOM_FUNCTION> (mapping = aes( <MAPPINGS> ),
 stat = <STAT> , position = <POSITION> ) +                 Not required, sensible defaults supplied
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

ggplot(data = mpg, **aes**(x = cty, y = hwy)) Begins a plot that you finish by adding layers to. Add one geom function per layer.

aesthetic mappings | data | geom

qplot(x = cty, y = hwy, data = mpg, geom = "point") Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

last_plot() Returns the last plot

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

```
a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))
```

a + geom_blank()
(Useful for expanding limits)

b + geom_curve(aes(yend = lat + 1, xend=long+1,curvature=z)) - x, xend, y, yend, alpha, angle, color, curvature, linetype, size

a + geom_path(lineend="butt", linejoin="round", linemitre=1) x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(group = group)) x, y, alpha, color, fill, group, linetype, size

b + geom_rect(aes(xmin = long, ymin=lat, xmax= long + 1, ymax = lat + 1)) - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin=unemploy - 900, ymax=unemploy + 900)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

b + geom_abline(aes(intercept=0, slope=1))
b + geom_hline(aes(yintercept = lat))
b + geom_vline(aes(xintercept = long))

b + geom_segment(aes(yend=lat+1, xend=long+1))
b + geom_spoke(aes(angle = 1:1155, radius = 1))

ONE VARIABLE continuous

```
c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)
```

c + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian") x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot() x, y, alpha, color, fill

c + geom_freqpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight

c2 + geom_qq(aes(sample = hwy)) x, y, alpha, color, fill, linetype, size, weight

discrete

```
d <- ggplot(mpg, aes(fl))
```

d + geom_bar()
x, alpha, color, fill, linetype, size, weight

TWO VARIABLES

continuous x , continuous y

```
e <- ggplot(mpg, aes(cty, hwy))
```

e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_jitter(height = 2, width = 2) x, y, alpha, color, fill, shape, size

e + geom_point(), x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl"), x, y, alpha, color, linetype, size

e + geom_smooth(method = lm), x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE), x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

discrete x , continuous y

```
f <- ggplot(mpg, aes(class, hwy))
```

f + geom_col(), x, y, alpha, color, fill, group, linetype, size

f + geom_boxplot(), x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight

f + geom_dotplot(binaxis = "y", stackdir = "center"), x, y, alpha, color, fill, group

f + geom_violin(scale = "area"), x, y, alpha, color, fill, group, linetype, size, weight

discrete x , discrete y

```
g <- ggplot(diamonds, aes(cut, color))
```

g + geom_count(), x, y, alpha, color, fill, shape, size, stroke

continuous bivariate distribution

```
h <- ggplot(diamonds, aes(carat, price))
```

h + geom_bin2d(binwidth = c(0.25, 500))
x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d()
x, y, alpha, colour, group, linetype, size

h + geom_hex()
x, y, alpha, colour, fill, size

continuous function

```
i <- ggplot(economics, aes(date, unemploy))
```

i + geom_area()
x, y, alpha, color, fill, linetype, size

i + geom_line()
x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv")
x, y, alpha, color, group, linetype, size

visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
j <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))
```

j + geom_crossbar(fatten = 2)
x, y, ymax, ymin, alpha, color, fill, group, linetype, size

j + geom_errorbar(), x, ymax, ymin, alpha, color, group, linetype, size, width (also **geom_errorbarh()**)

j + geom_linerange()
x, ymin, ymax, alpha, color, group, linetype, size

j + geom_pointrange()
x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

maps

```
data <- data.frame(murder = USArrests$Murder,
state = tolower(rownames(USArrests)))
map <- map_data("state")
k <- ggplot(data, aes(fill = murder))
```

k + geom_map(aes(map_id = state), map = map) **+ expand_limits**(x = map$long, y = map$lat), map_id, alpha, color, fill, linetype, size

THREE VARIABLES

```
seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))l <- ggplot(seals, aes(long, lat))
```

l + geom_contour(aes(z = z))
x, y, z, alpha, colour, group, linetype, size, weight

l + geom_raster(aes(fill = z), hjust=0.5, vjust=0.5, interpolate=FALSE)
x, y, alpha, fill

l + geom_tile(aes(fill = z)), x, y, alpha, color, fill, linetype, size, width

Revisão

Variáveis hidrológicas e de qualidade da água têm características inerentes aos seus processos naturais e/ou seus métodos de medição

- auxiliam na Análise Exploratória de Dados
- promovem a identificação de inconsistências
- balizam a escolha dos métodos estatísticos

A representação gráfica dos dados é de suma importância para a compreensão e para a comunicação das informações contidas em uma amostra

- cada gráfico tem um propósito que deve condizer com o que se quer mostrar
- cada representação tem vantagens e limitações
- a experiência ajuda (e muito!) na escolha

Para a próxima aula

Pesquisar ao menos um artigo científico que mostre representações gráficas úteis (ou não!!) das amostras em análise

A escolha do artigo, revista e tema é livre. Recomenda-se preferência por trabalhos do estado-da-arte (publicações nos últimos 3 a 5 anos)

Preparar uma apresentação (5 min.) para explicar a representação gráfica escolhida

- não é preciso estudar o artigo completo! Foco na representação gráfica!
- a atividade pode ser feita em duplas

Estatística Aplicada a Ciências Ambientais

Daniel Detzel
detzel@ufpr.br