SME0803 Visualização e Exploração de Dados

Representação de dados quantitativos

Prof. Cibele Russo

cibele@icmc.usp.br

Baseado em

Murteira, B. J. F., Análise Exploratória de Dados. McGraw-Hill, Lisboa, 1993.

Mário de Castro. Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

Variáveis discretas

Dados: n observações de uma variável discreta x.

Existem m diferentes valores $x_1 < x_2 < .... < x_m$, $1 \leq m \leq n$.

Tabela de frequências: tabela com os valores de xj e uma das ou ambas as frequencias f_j e $f_j^\star$, j = 1,..,m.

x	frequência absoluta	frequência relativa
x_1	f_1	$f_1^\star$
x_2	f_2	$f_2^\star$
$\vdots$	$\vdots$	$\vdots$
x_m	f_m	$f_m^\star$
Total	n	1(100%)

As frequências acumuladas F_j e $F_j^\star$ estão bem definidas, $j = 1, ..., m$ e podem ser uma coluna de uma tabela de frequências.

Gráfico de linhas verticais

Representação em linhas verticais das frequências absolutas

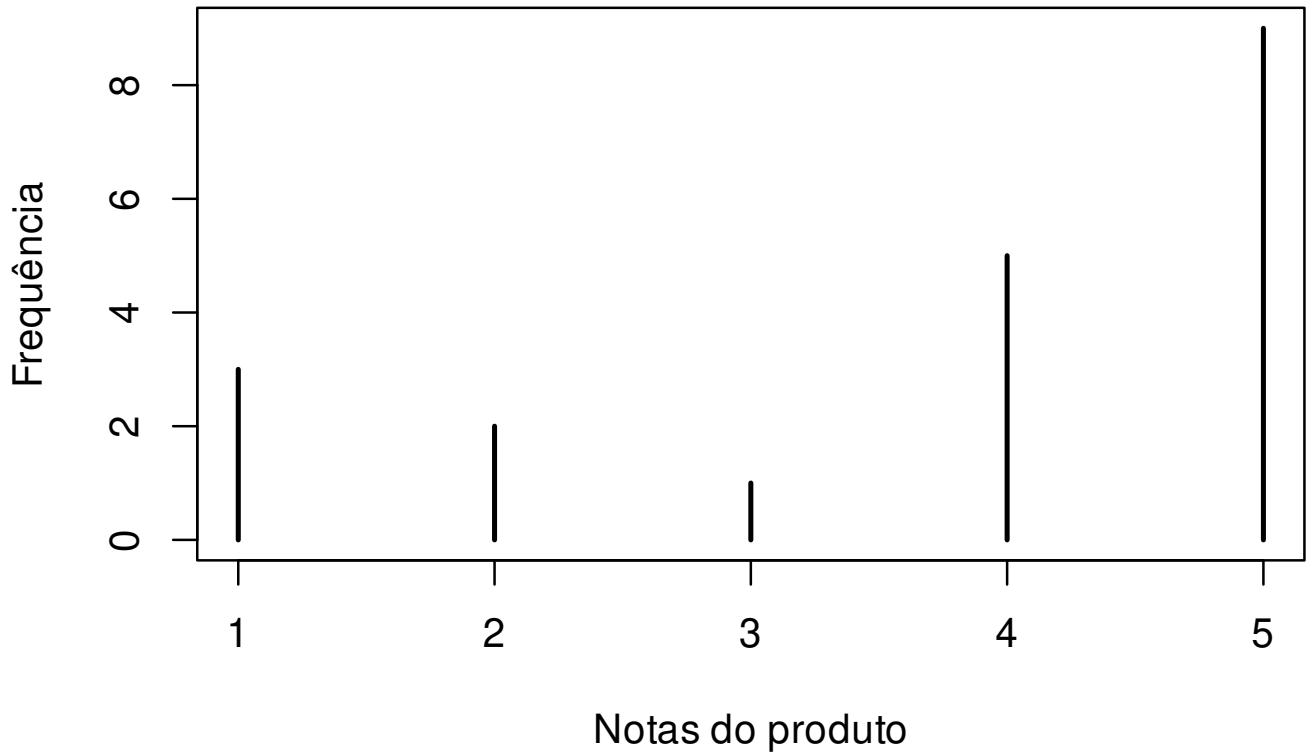


Gráfico de barras

Representação das frequências absolutas em retângulos.

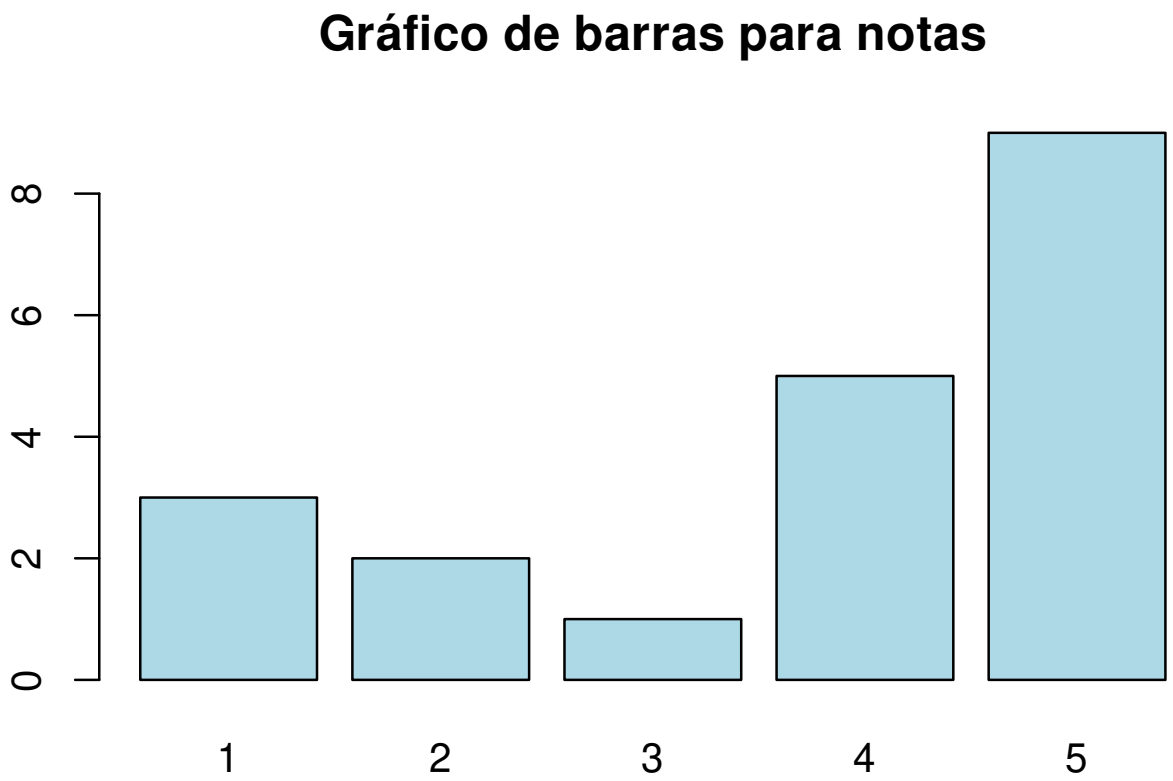


Gráfico de barras

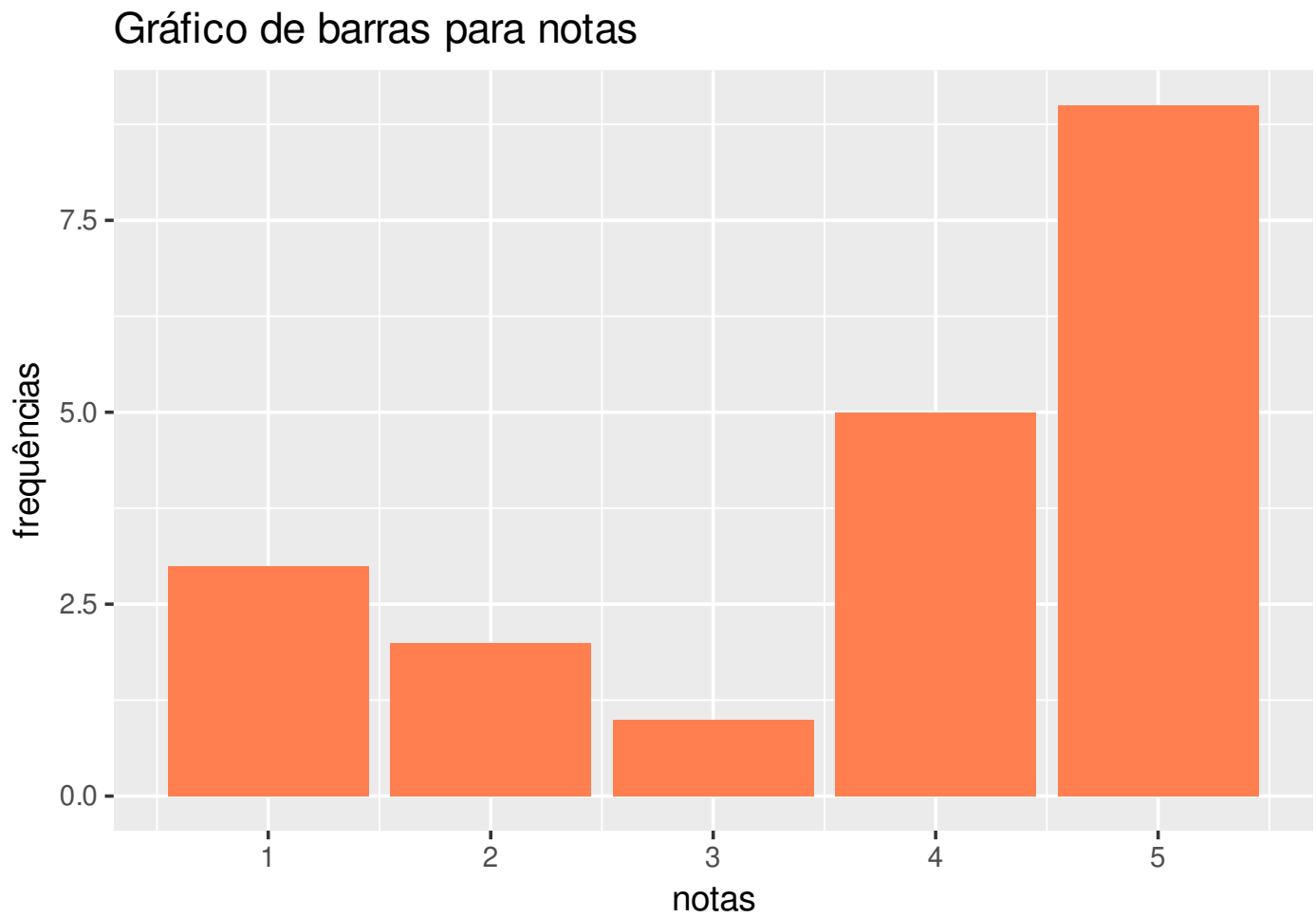
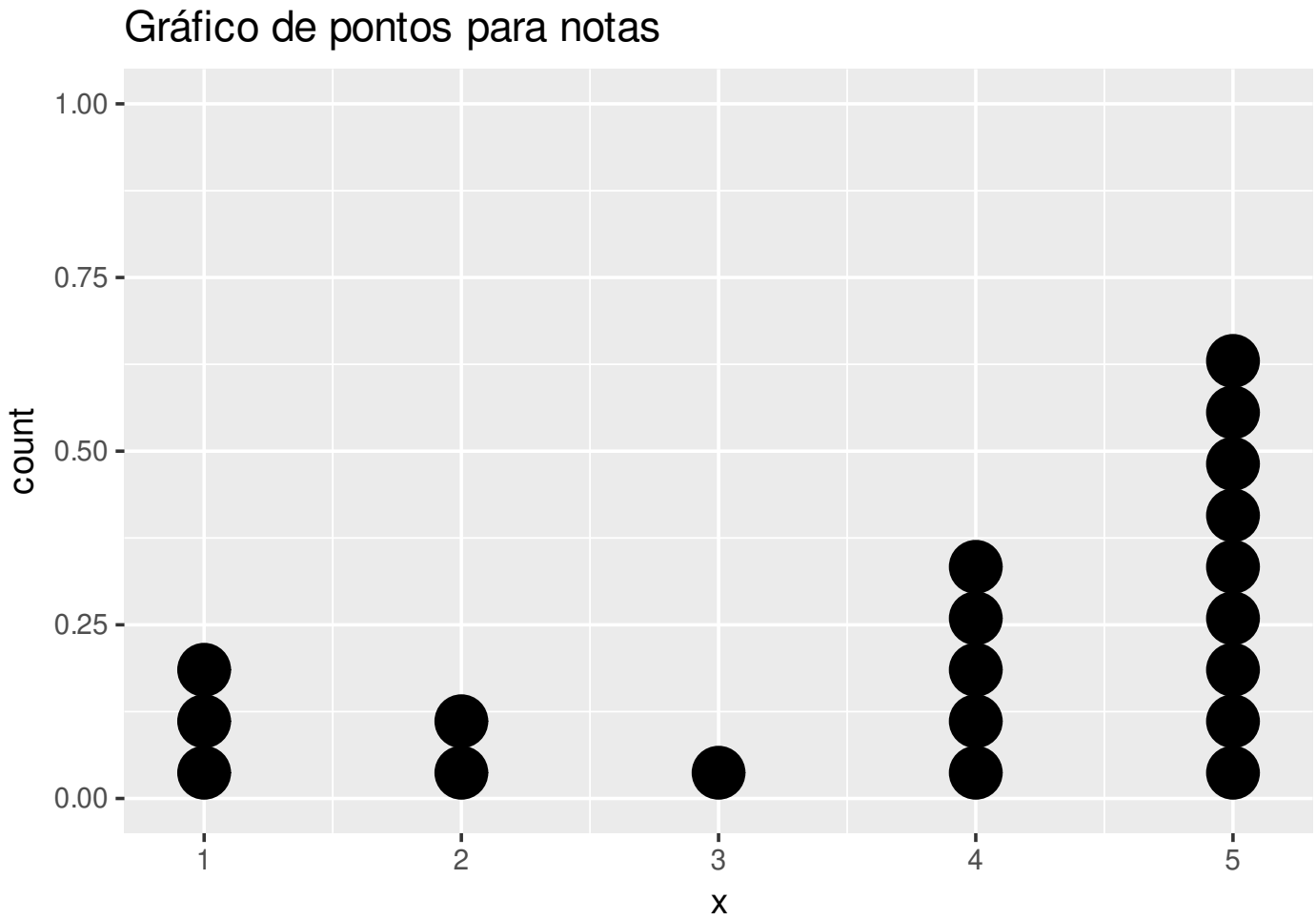


Gráfico de pontos

Cada observação é representada por um ponto. Valores repetidos produzem pontos empilhados.



Variáveis contínuas

Dados: n observações de uma variável contínua x.

Existem m diferentes valores $x_1 < x_2 < ... < x_m, 1 \leq m \leq n$.

Tabela de frequências. Se m é "grande", uma tabela com todos os diferentes valores não cumpre o papel de resumir os dados.

Variáveis contínuas

Representação em k **intervalos de classe** (ou **classes**) do tipo $[LIj, LSj),\ j = 1, ..., k$.

LIj: limite inferior e LSj: limite superior.

Construção:

1. Escolha do número de classes (k). Usualmente, $5 \leq k \leq 15$.
2. Cálculo da amplitude (A) : $A = max - min$, sendo que min e max são o menor e o maior valor dos dados.
3. Cálculo da amplitude de classe (h) : $h = A/k$.
4. Obtenção dos limites das classes: $LI1 = min$, $LS1 = LI1 + h$, $LI2 = LS1$, $LS2 = LI2 + h$, ..., $LIk = LSk - 1$, $LSk = max$.

Variáveis contínuas

Observações:

1. h e $LI1$ podem ser arredondados por conveniência.
2. Cada valor observado de x pertence a uma e apenas uma classe.
3. h pode variar com a classe.

Variáveis contínuas

Ponto médio da classe (ou marca de classe): $X_j = \frac{LI_j + LS_j}{2}$

Frequência absoluta da classe (f_j): número de observações $\in [LI_j, LS_j)$).

Frequência relativa de cada intervalo de classe: $f_j^\star = f_j/n$.

Frequência acumulada da classe (F_j):

$$F_j = f_1 + f_2 + \ldots + f_j = \sum_{i=1}^{j} f_i, \; F_k = n$$

Frequência acumulada relativa da classe:

$$F_j^\star = \frac{F_j}{n}, F_k^\star = 1$$

Obs. Na representação por classes há perda de informação.

Histograma

Densidade de frequência (ou densidade):

$$f_{d_j} = \frac{f_j}{h_j} \text{ ou } f^{\star}_{d_j} = \frac{f_j\star}{h_j}, j = 1, \ldots, k$$

Representação gráfica:

Histograma (histogram - Karl Pearson, 1895):

Gráfico de barras adjacentes com bases iguais às amplitudes das classes e alturas iguais às densidades.

Obs. Se as classes tiverem amplitude constante, as alturas das barras usualmente são iguais às frequencias.

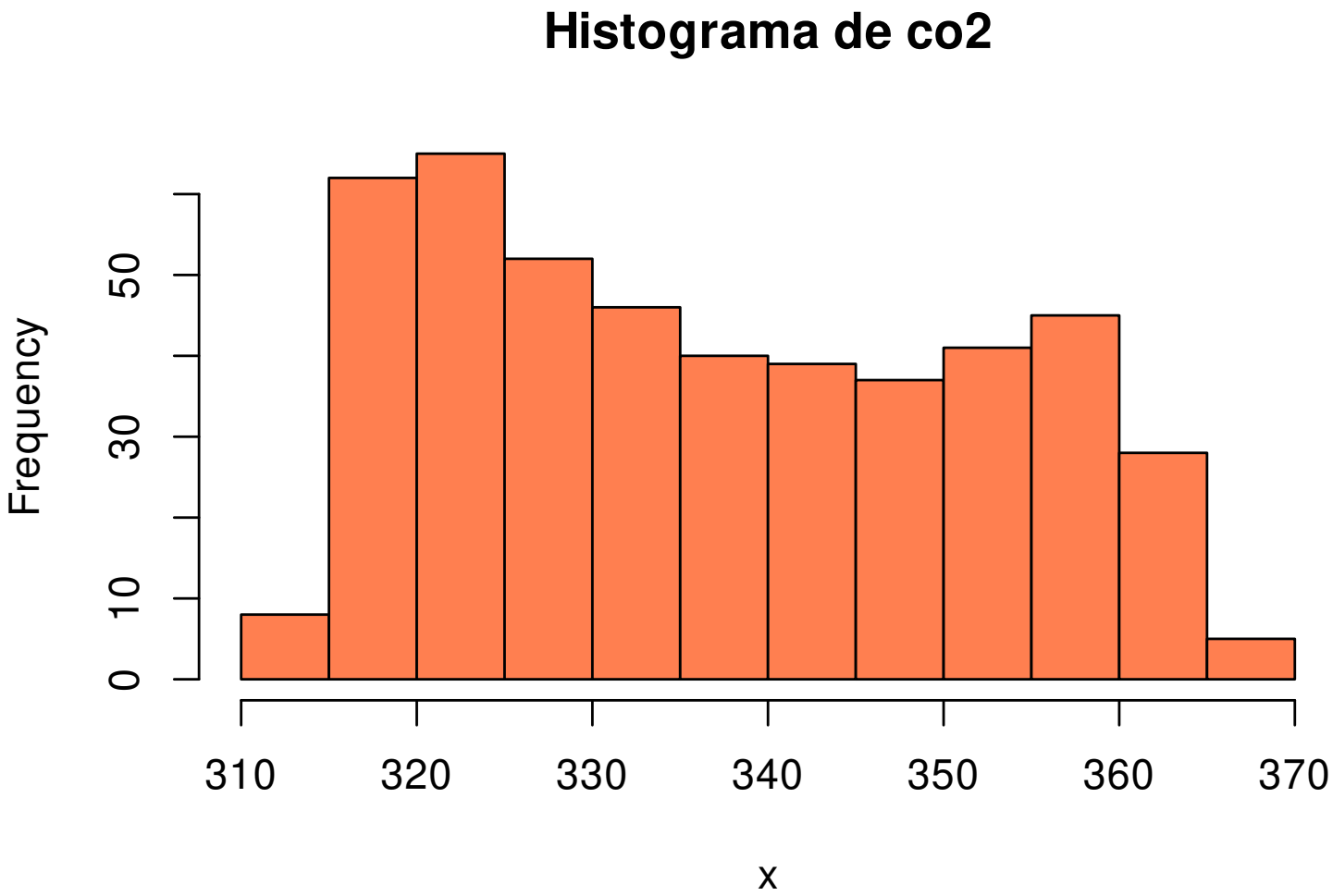
Histograma

Propriedades do histograma:

- $\sum_{j=1}^{k} h_j f_{d_j} = \sum_{j=1}^{k} h_j \frac{f_j}{h_j} = \sum_{j=1}^{k} f_j = n$
- $\sum_{j=1}^{k} h_j f^{\star}_{d_j} = \sum_{j=1}^{k} h_j \frac{f^{\star}_j}{h_j} = \sum_{j=1}^{k} f^{\star}_j = 1$

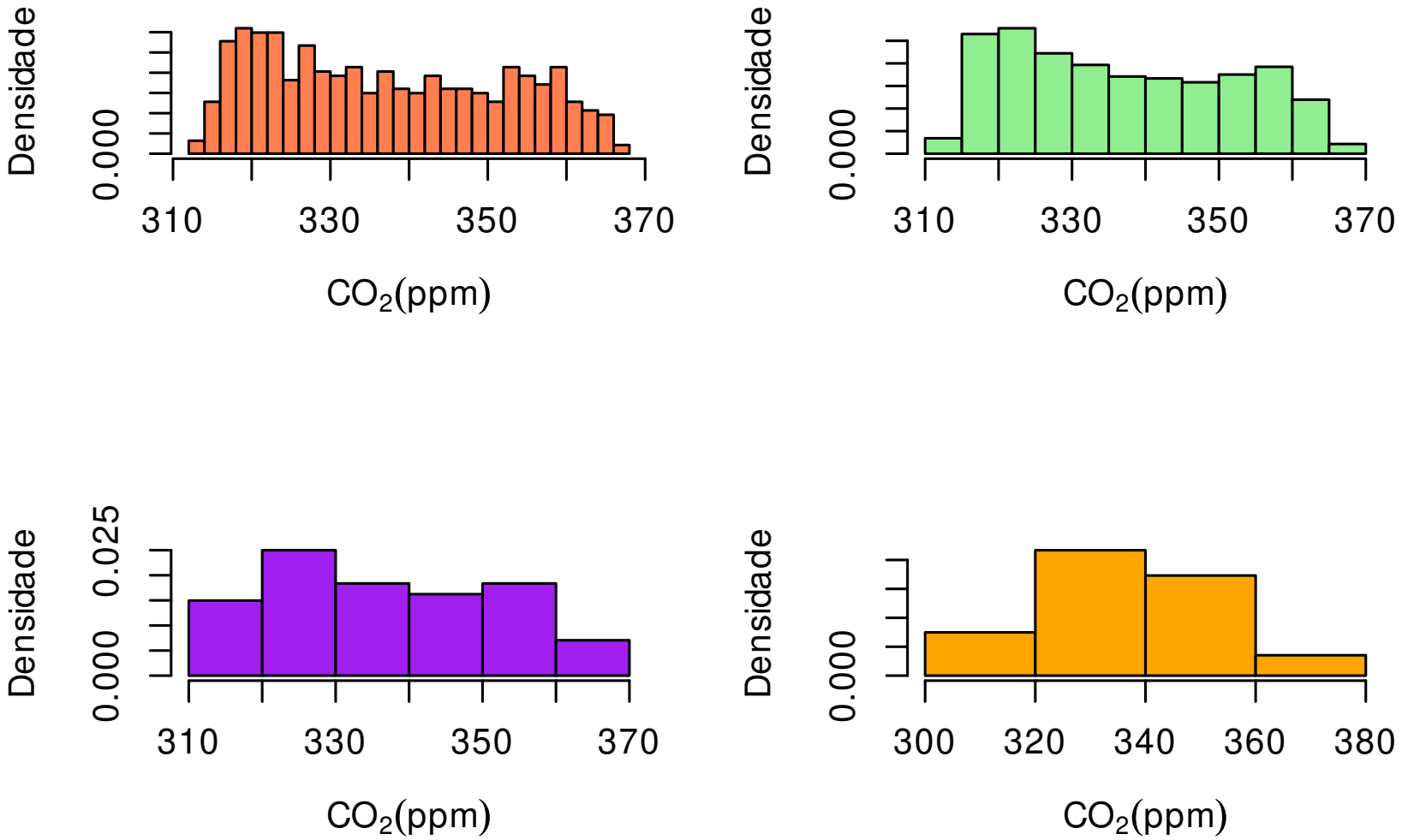
Obs. Na construção de um histograma, quanto maior for n, melhor.

Histograma



O histograma fornece uma ideia da distribuição dos dados.

Variáveis contínuas



Diferentes números de classes

Variáveis contínuas

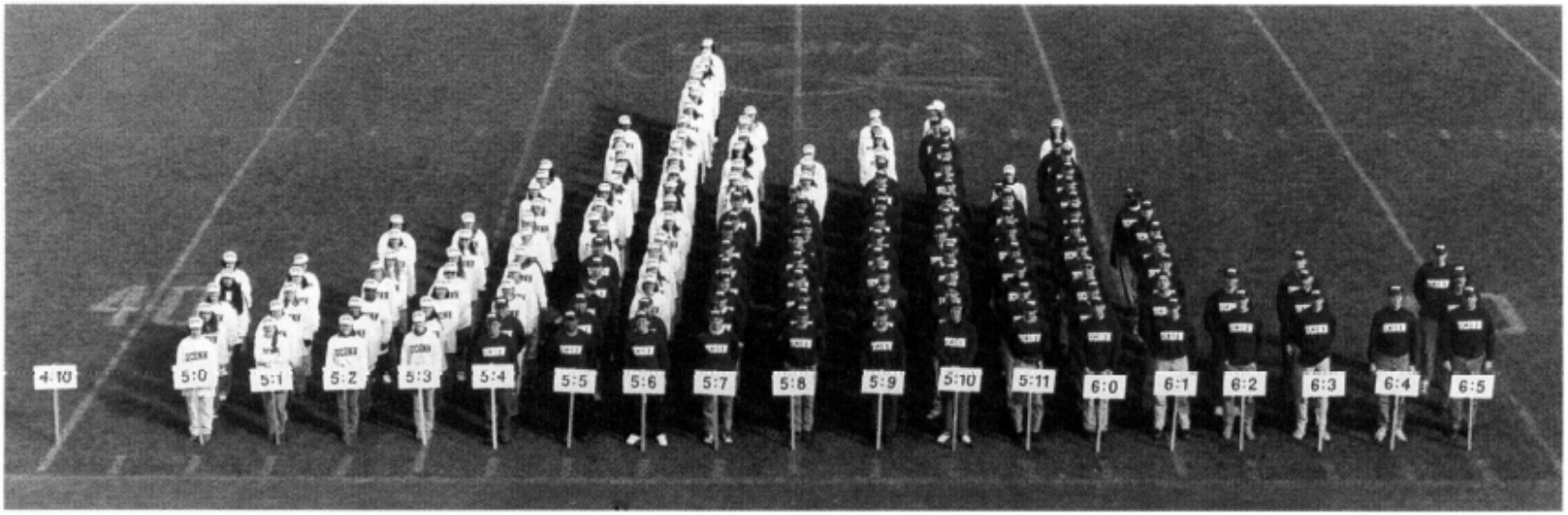


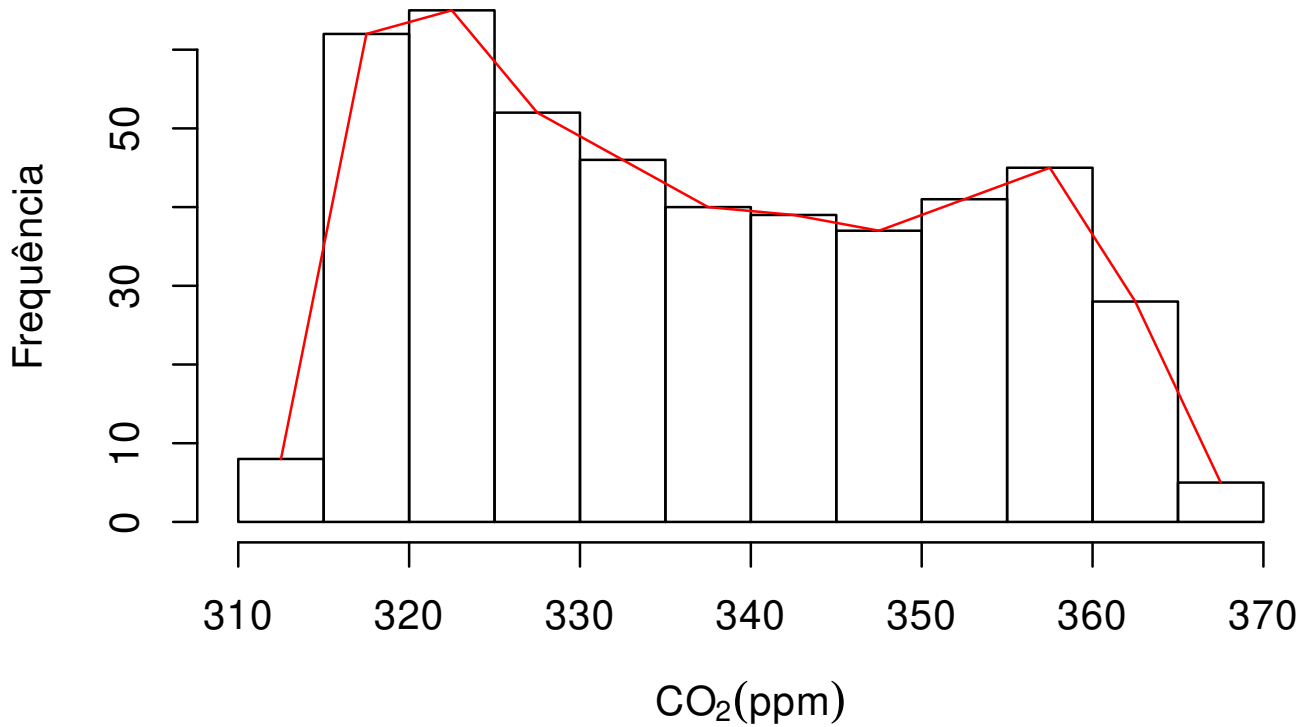
Figure 7. Living histogram of 143 student heights at University of Connecticut.

Histograma humano.

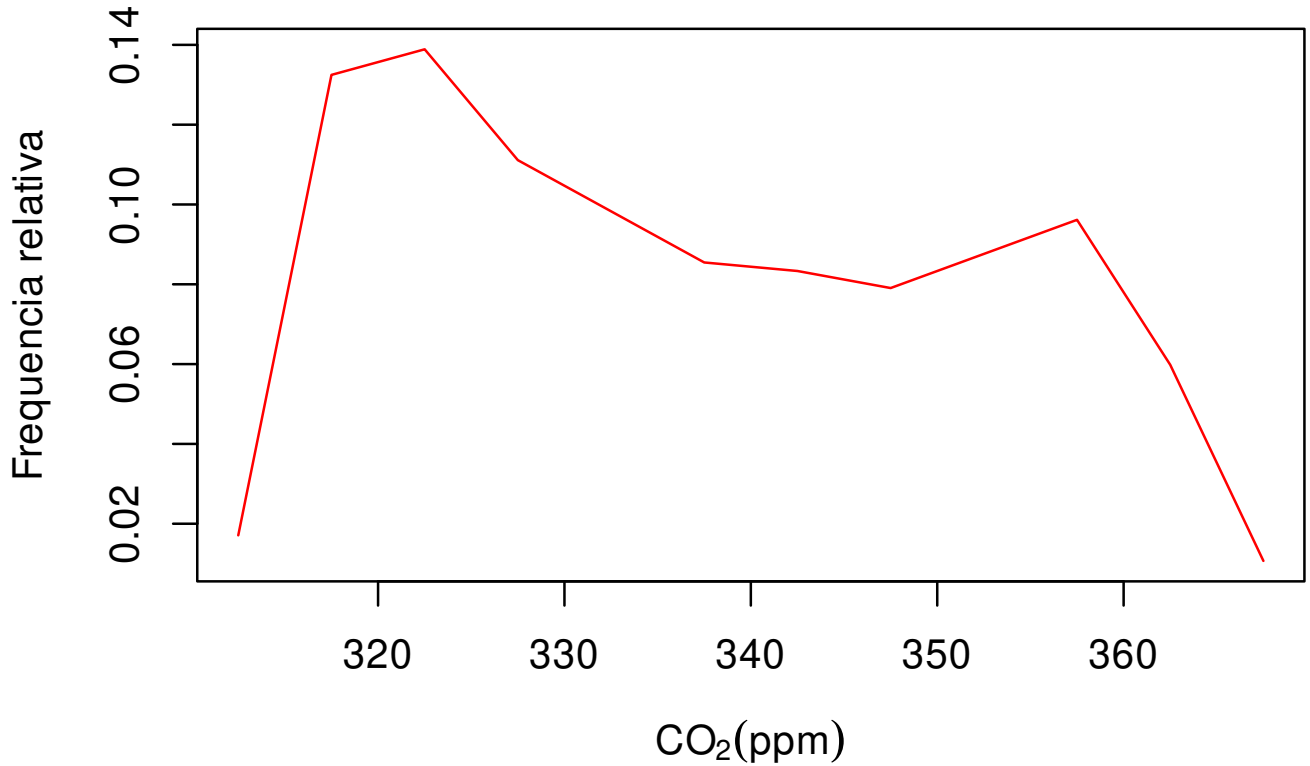
Fonte: The American Statistican 56(3), 223 - 229, 2002.

Polígono de frequencias

Formado pelos segmentos unindo os pontos centrais dos topos das barras.

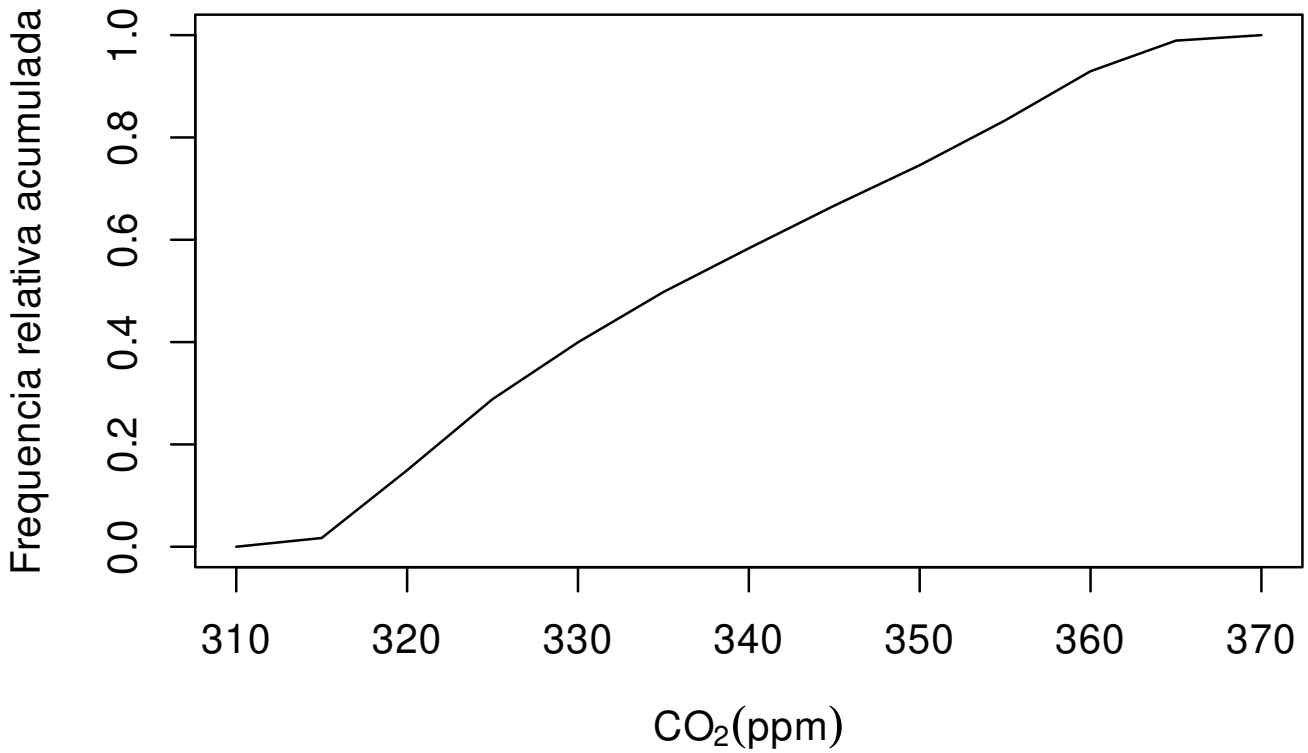


Polígono de frequências



Polígono de frequencias acumuladas (ogiva)

Formado por segmentos de retas unindo o limite superior das classes no topo das barras.



Polígono de frequencias acumuladas (ogiva)

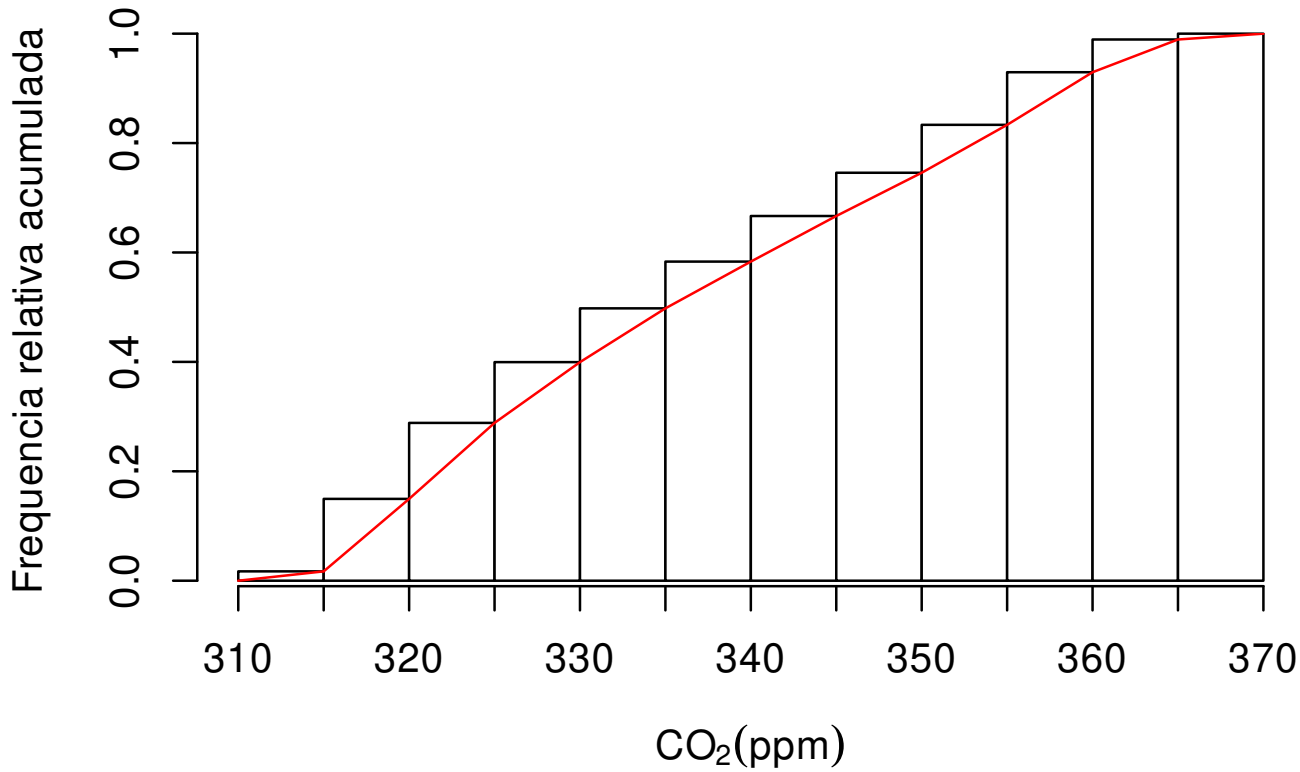
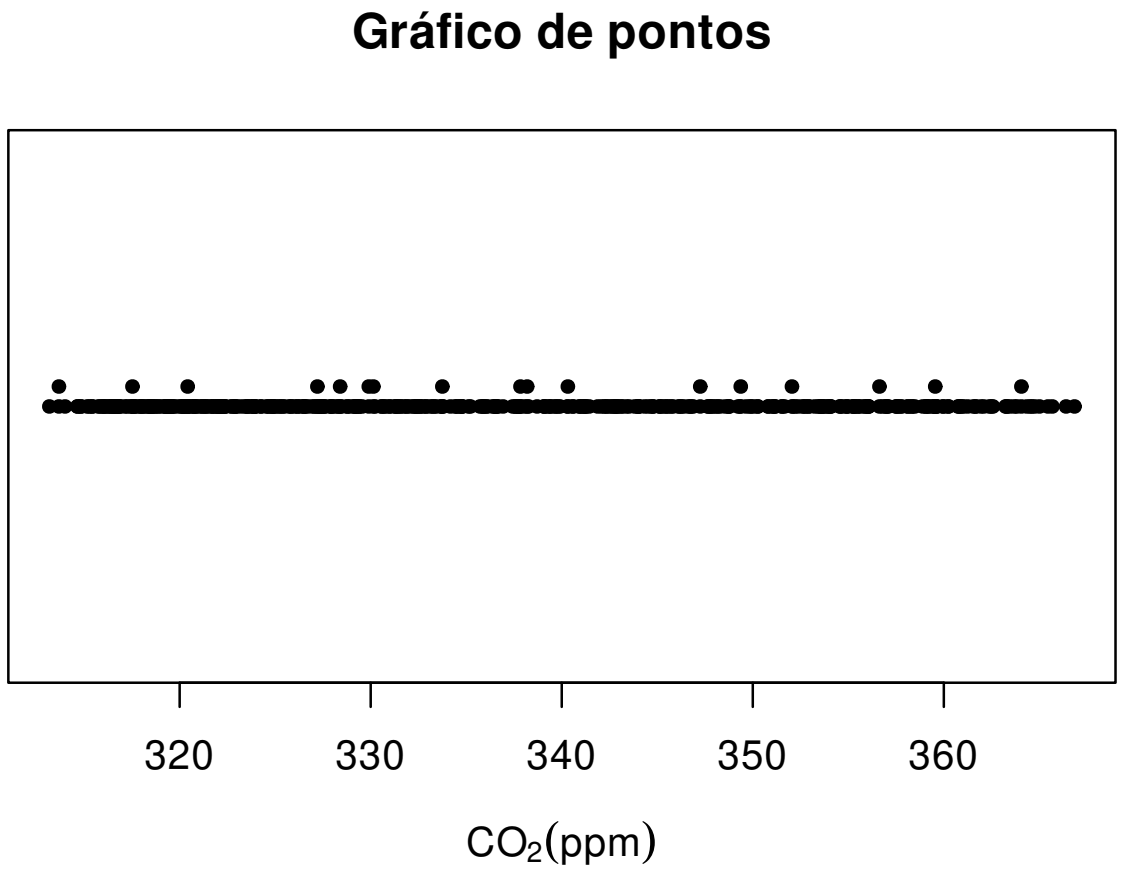


Gráfico de pontos

Cada observação é representada por um ponto.

Não há perda de informação.

Se n for grande, o gráfico pode perder em clareza.



[]

Gráfico de pontos

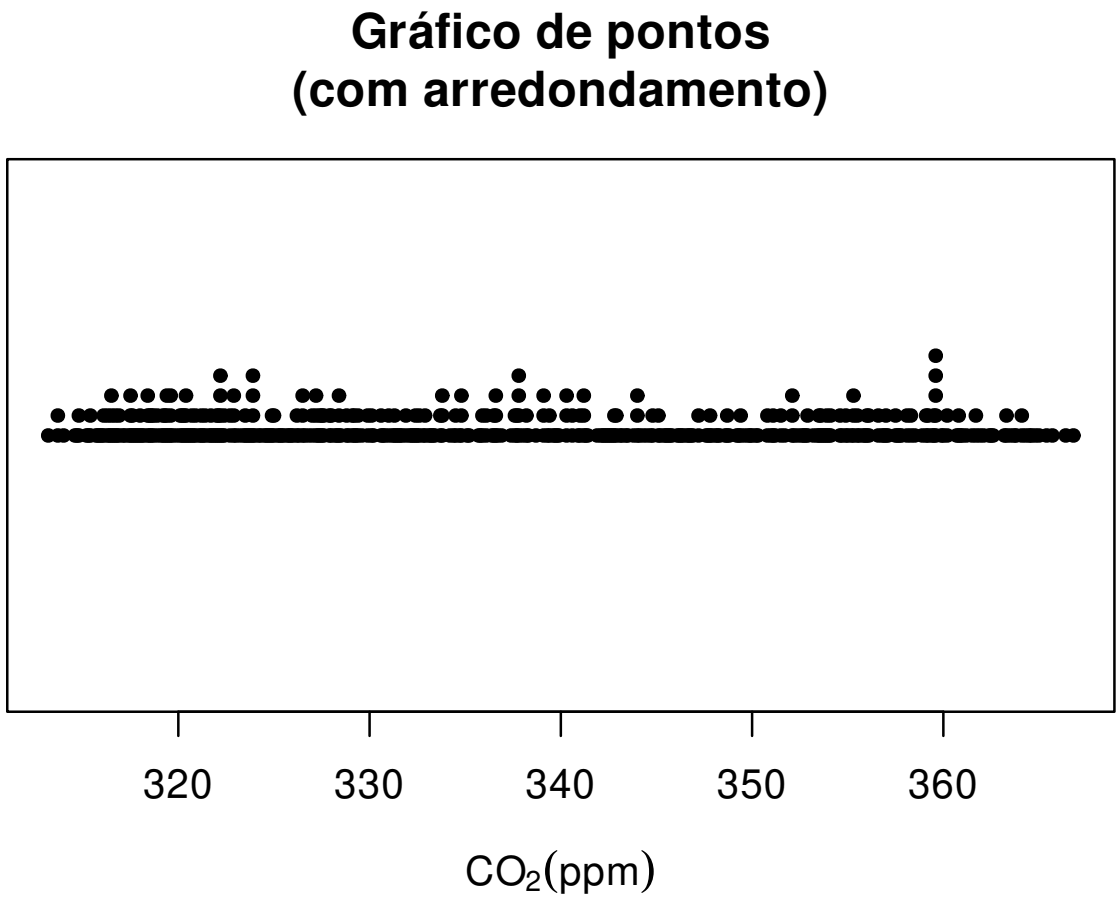
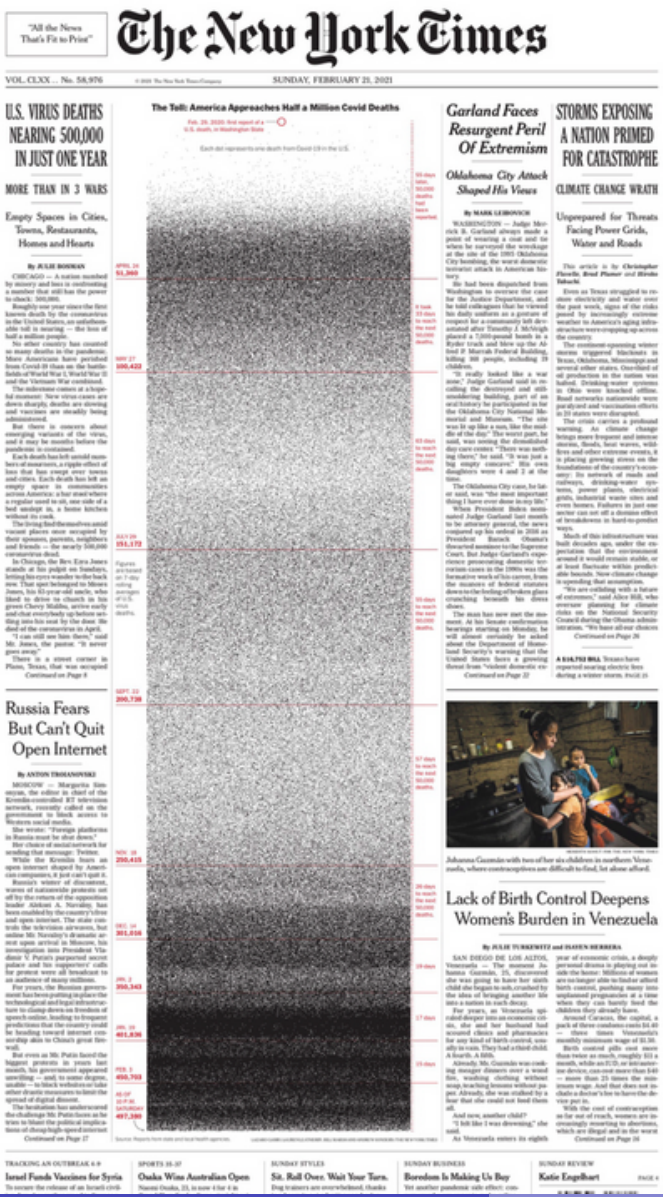


Gráfico de pontos


"All the News That's Fit to Print"

The New York Times

VOL. CLXX .. No. 58,976 SUNDAY, FEBRUARY 21, 2021

U.S. VIRUS DEATHS NEARING 500,000 IN JUST ONE YEAR

MORE THAN IN 3 WARS

Empty Spaces in Cities, Towns, Restaurants, Homes and Hearts

The Toll: America Approaches Half a Million Covid Deaths

Garland Faces Resurgent Peril Of Extremism

Oklahoma City Attack Shaped His Views

STORMS EXPOSING A NATION PRIMED FOR CATASTROPHE

CLIMATE CHANGE WRATH

Unprepared for Threats Facing Power Grids, Water and Roads

Russia Fears But Can't Quit Open Internet

Lack of Birth Control Deepens Women's Burden in Venezuela

New York Times em 21/02/2021:

https://pbs.twimg.com/media/EuwfGryXAAE6zhc?format=jpg&name=large

Gráfico de linhas

Representa variáveis coletadas com referência a uma unidade de tempo. Chamadas de séries históricas ou séries temporais (time series). Séries temporais podem ser de variáveis discretas ou qualitativas.

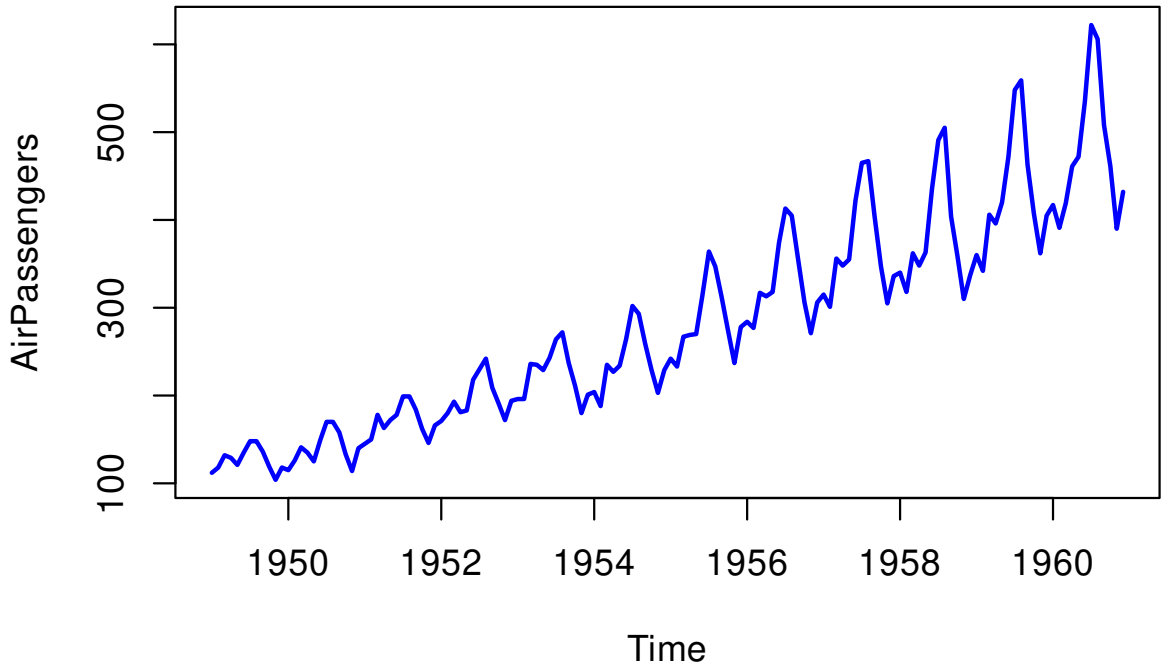


Gráfico de linhas

Gráfico de linhas do índice de Gini de 2012 a 2019.

Fonte: `https://portal.fgv.br/noticias/desigualdade-renda-brasil-bate-recorde-aponta-levantar`

Gráfico de ramos-e-folhas

Para um conjunto de dados quantitativos contínuos, considerando algum arredondamento específico, representamos a parte fracionária com **folhas** e a parte inteira são os **ramos**.

Motivação: Considere as notas de 100 alunos em uma prova.

```
> stem(notas)

  The decimal point is at the |

   1 | 8
   2 | 56
   3 | 14444556789
   4 | 001112244456666677889
   5 | 00111233333455566778999
   6 | 000122445568889
   7 | 00122236777899
   8 | 234579
   9 | 00336
  10 | 0
```

Fonte: adaptado das notas de aula de M. Castro.

Gráfico de ramos-e-folhas (stem-and-leaf display ou stemplot)

Representação com **nenhuma ou pouca perda de informação**.

Cada valor observado da variável é dividido em duas partes: **ramo** (dígitos dominantes) e **folha** (dígitos dominados).

Os ramos se situam à esquerda de uma linha vertical e as folhas à direita. O número de ramos é escolhido.

Usualmente uma folha representa o último dígito de um número (números podem ser arredondados ou representados como múltiplos de potências de 10).

Os dígitos restantes de um número compõem o ramo.

Gráfico de caixa (boxplot)

Representação gráfica inteligente que permite a observação da **localização**, **dispersão**, **assimetria**, **pontos discrepantes** (outliers).

Além disso, permite comparar visualmente a distribuição de dados em dois grupos. Pode indicar evidências sobre a igualdade das médias entre os dados de dois grupos, pendente de análise confirmatória inferencial.

Gráfico de caixa (boxplot)

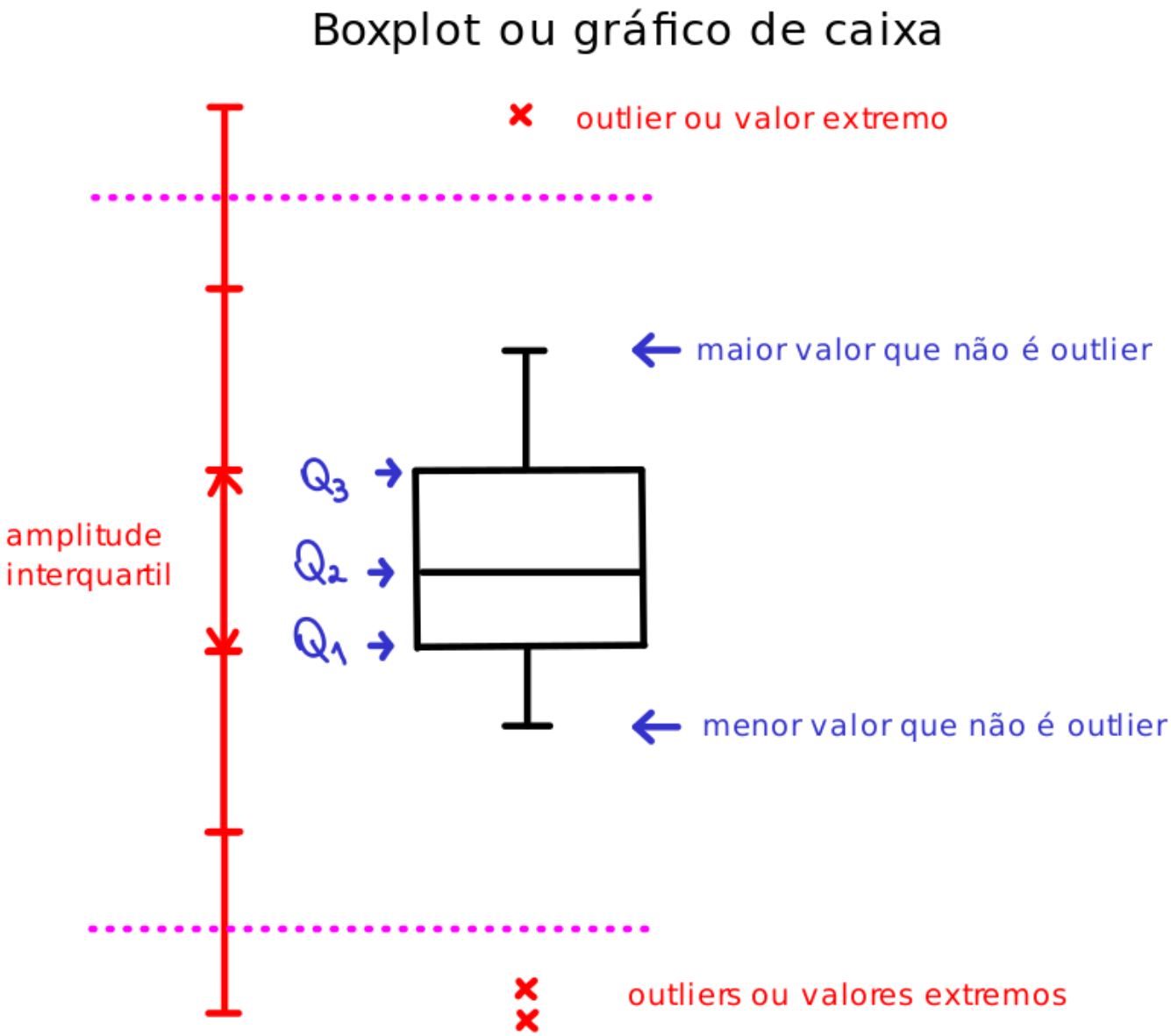
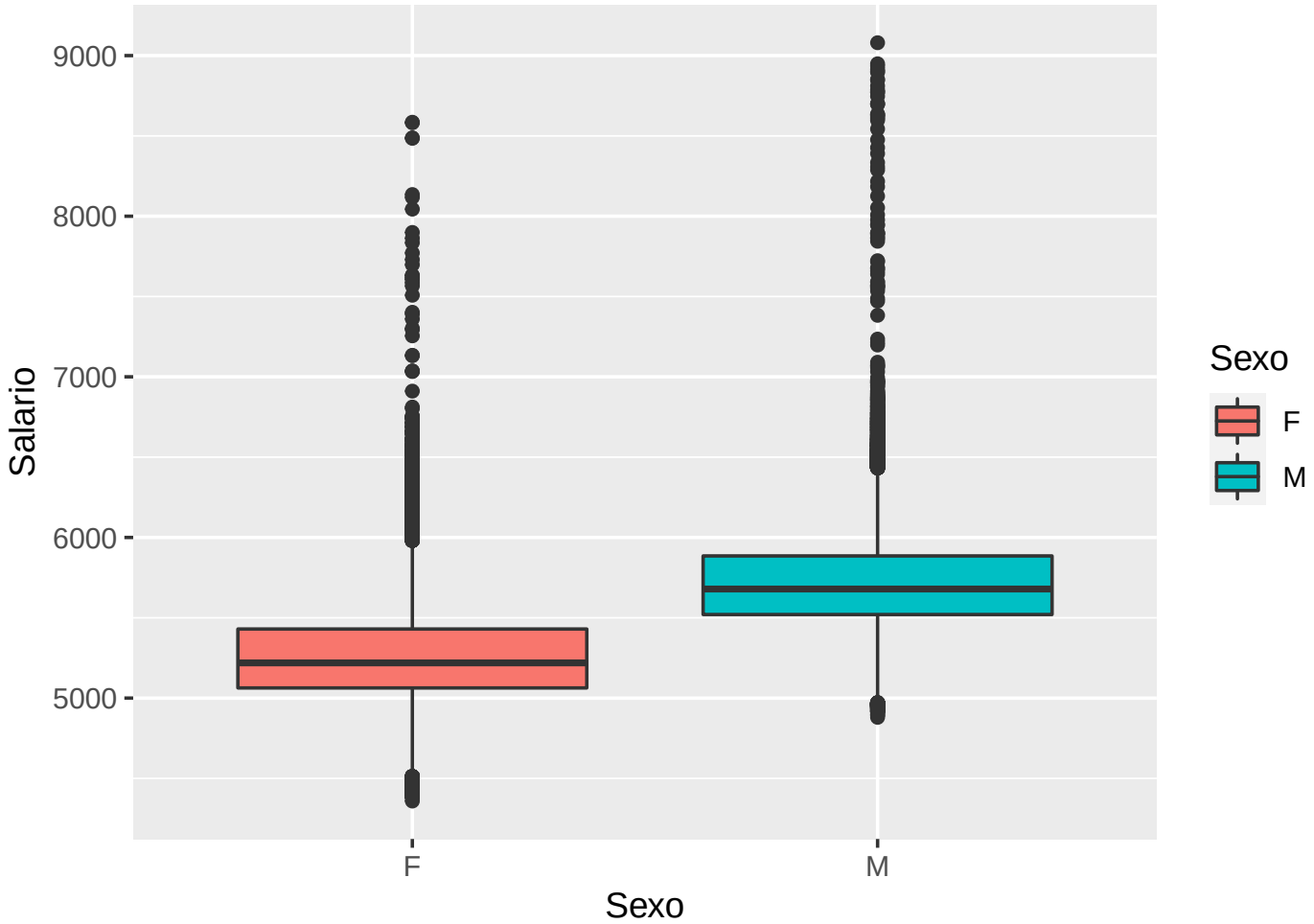
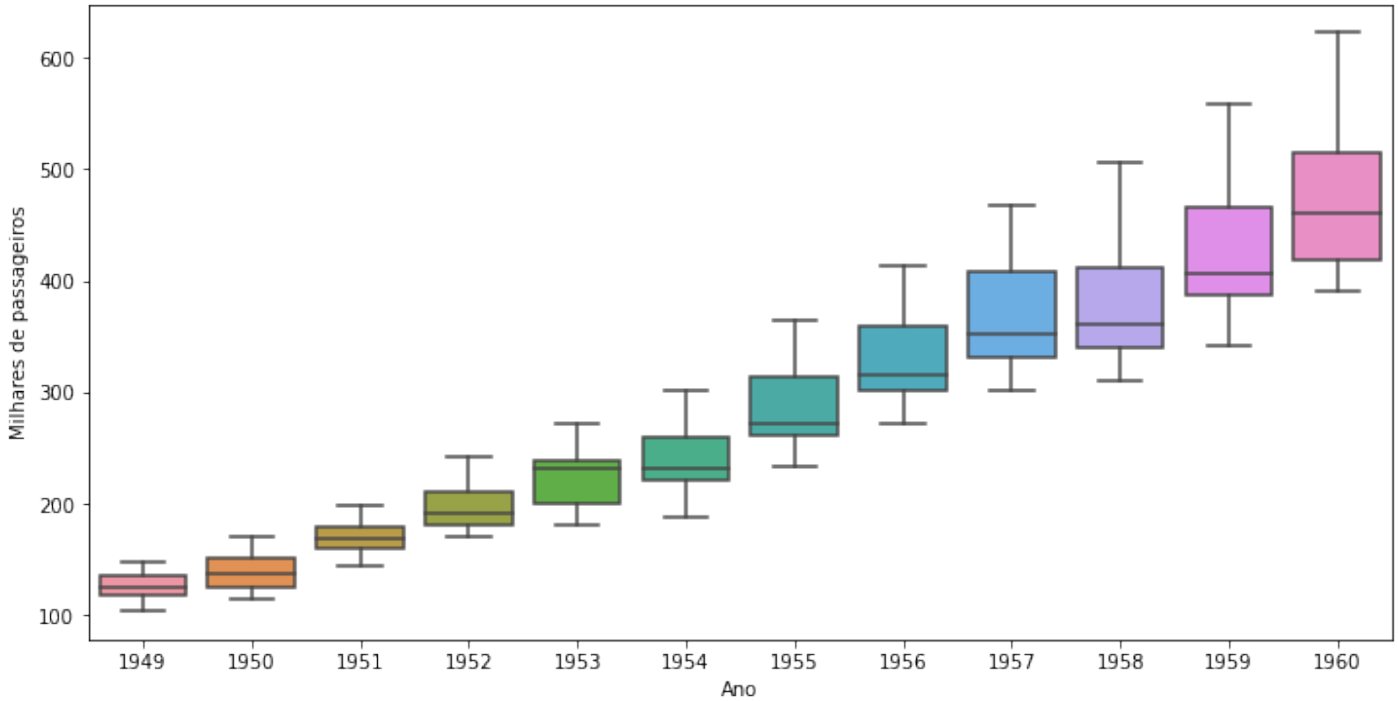


Gráfico de caixa (boxplot)



Boxplots de salários por sexo.

Gráfico de caixa (boxplot)



Boxplots de dados anuais de passageiros aéreos.