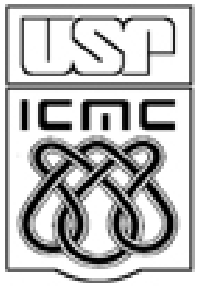


3. Representação de variáveis quantitativas

2010

3.1 Variáveis discretas

Dados: n observações de uma variável discreta *x*.

Existem k diferentes valores $x_1 < x_2 < .... < x_k$, $1 \leq k \leq n$.

Tabela de frequencias: tabela com os valores de x_j e uma das ou ambas as frequencias f_j e f_j^*, j = 1,..,k.

Tabela 1. Descrição da tabela.

x	Frequencia	Frequencia relativa
x_1	f_1	f^*_1
x_2	f_2	f^*_2
...	...	...
x_k	f_k	f^*_k
Total	n	1 (100%)

As frequencias acumuladas F_j e F_j^* estão bem definidas, j = 1,...,k e podem ser uma coluna de uma tabela de frequencias.

Tabelas e gráficos em R

```
> x = c(2, 3, 3, 1, 0, 0, 2, 2, 2, 2, 2, 1, 2, 3, 2, 0, 2, 0, 2, 2, 1, 3, 1, 3, 5, 0, 3, 2, 3, 2, 2, 3, 1, 3, 3, 0, 2, 2, 2,  2)
```

```
> (n = length(x))
```

```
40
```

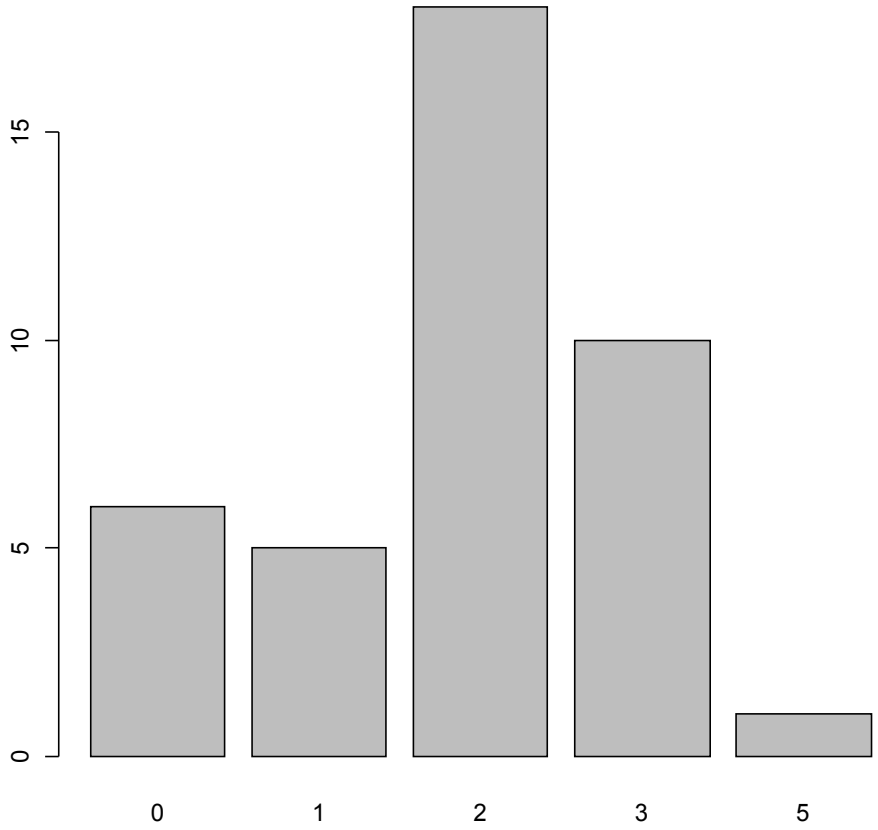
```
> (freqa = table(x))
```

```
 0  1  2  3  5
 6  5 18 10  1
```

```
> freqa / n * 100
```

```
    0     1     2     3     5
 15.0  12.5  45.0  25.0   2.5
```

```
> barplot(freqa)
```



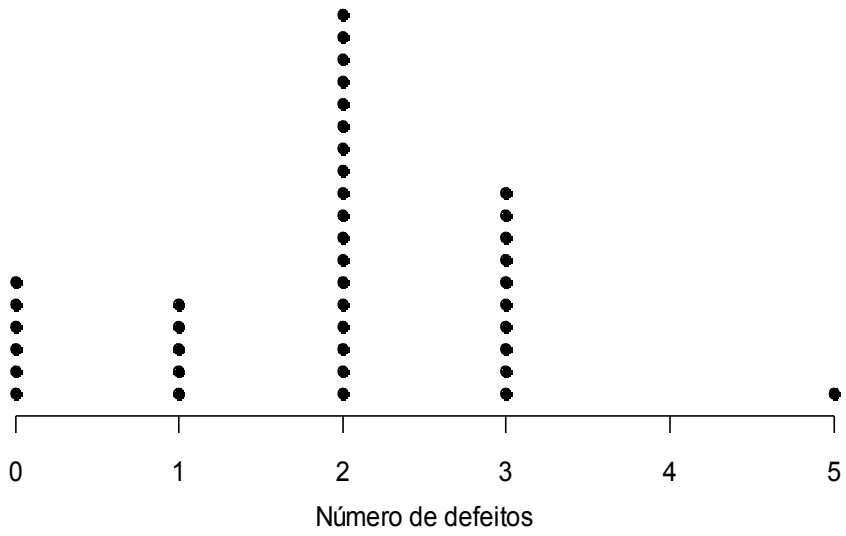
Tabelas e gráficos em R

Gráfico de pontos (*dot plot*)

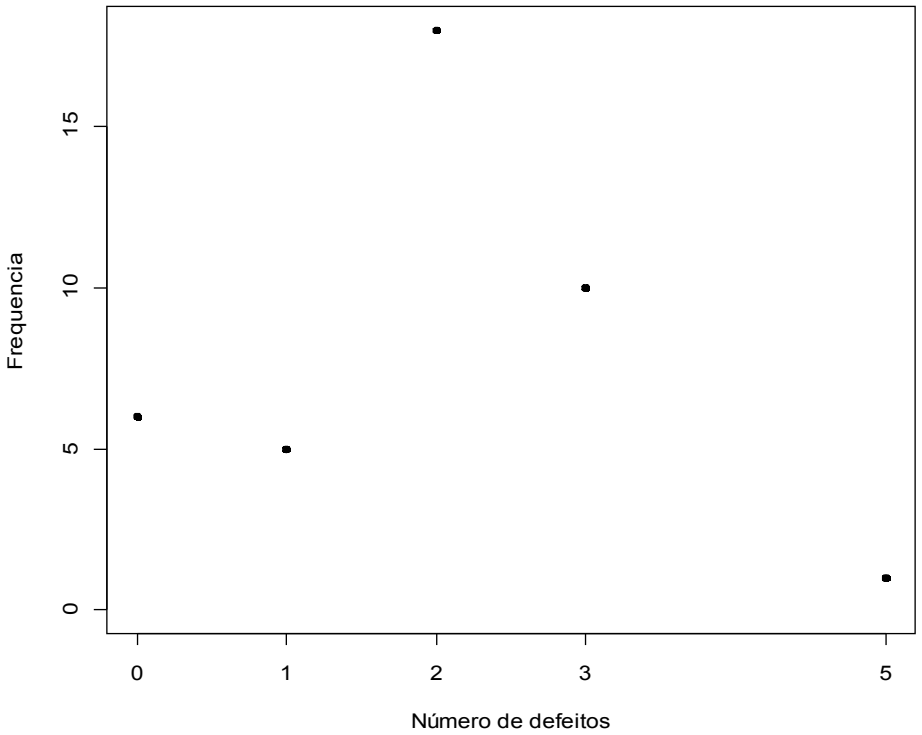
Cada observação é representada por um ponto. Valores repetidos produzem pontos empilhados.

> library(plotrix)

> dotplot.mtb(x, xlab = "Número de defeitos")

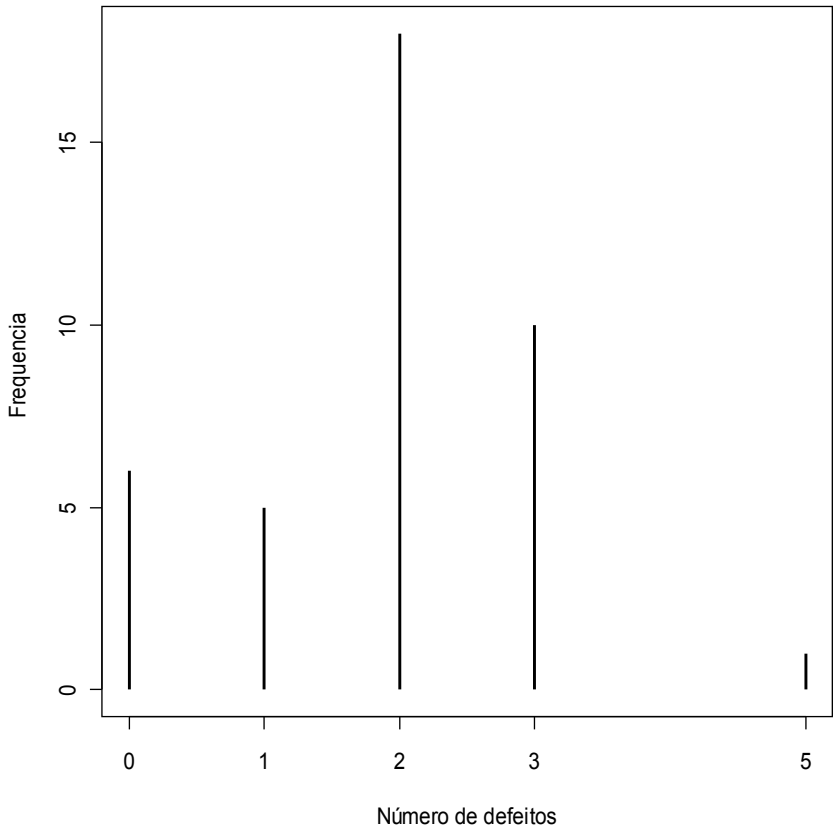


> plot(freqa, type = "p", pch = 20, xlab = "Número de defeitos", ylab = "Frequencia")



Tabelas e gráficos em R

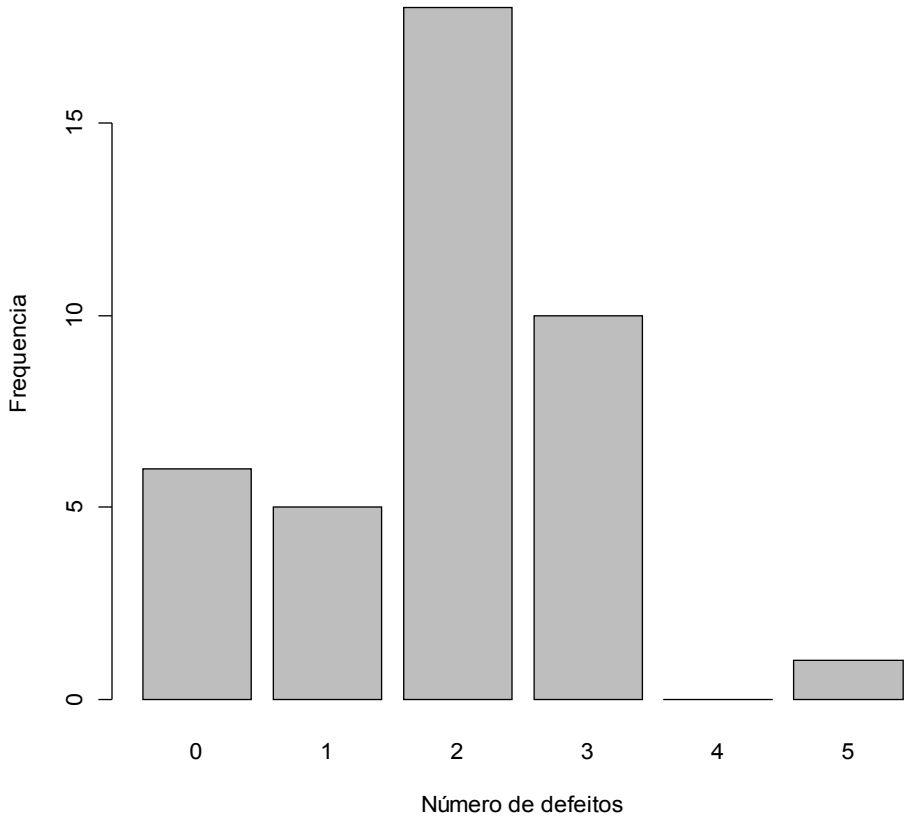
```
> plot(freqa, xlab = "Número de defeitos",
ylab = "Frequencia")
```




```
> freqaux = table(c(x, 4))

> freqaux[which(names(freqaux) == "4")] = 0

> barplot(freqaux, xlab = "Número de defeitos",
ylab = "Frequencia")
```



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

Representação em k intervalos de classe (ou classes) do tipo $[LI_j, LS_j)$, $j = 1,...,k$.

Li_j : limite inferior e LS_j : 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: $LI_1 = min$, $LS_1 = LI_1 + h$, $LI_2 = LS_1$, $LS_2 = LI_2 + h$, ..., $LI_k = LS_{k-1}$, $LS_k = MAX$.

Obs. (1) h e LI_1 podem ser arredondados por conveniência.

(2) Cada valor de x pertence a uma e apenas uma classe.

(3) h pode variar com a classe.

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

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

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

Frequência acumulada da classe (F_j):

$$F_j = f_1 + f_2 + \cdots + f_j = \sum_{l=1}^{j} f_l \qquad (F_k = n).$$

Frequência acumulada relativa da classe:

$$F_j^* = \frac{F_j}{n} \quad (F_k^* = 1).$$

Obs. Na representação por classes há perda de informação.

Densidade de frequência (ou densidade):

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

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.

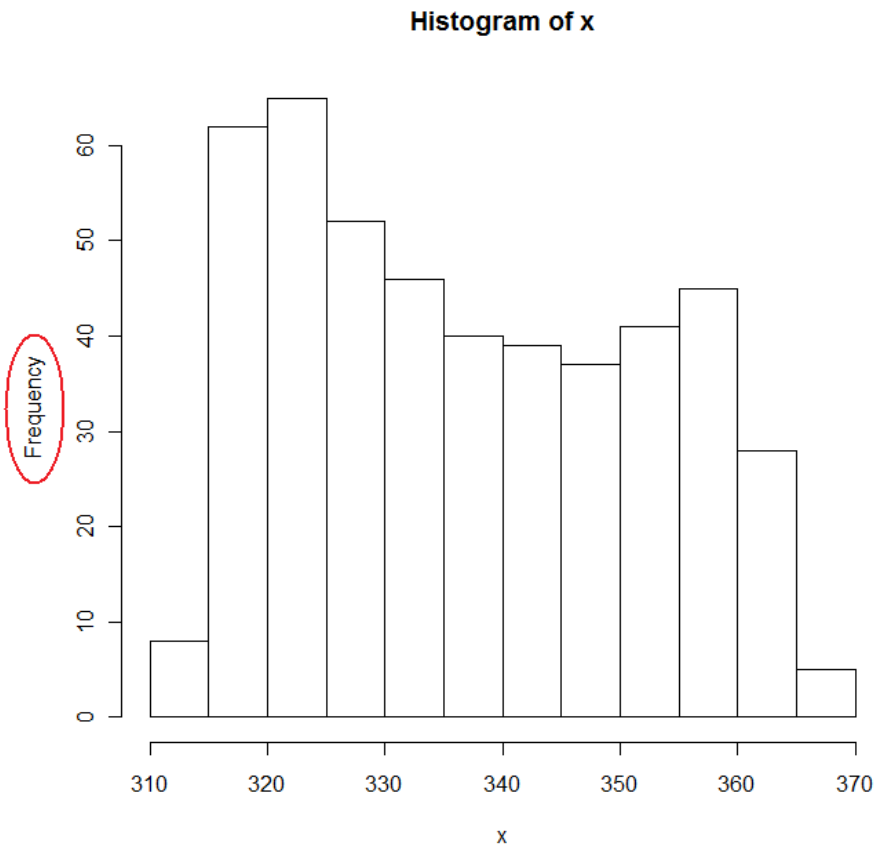
Propriedade :

$$\sum_{j=1}^{k} h_j f_{d_j} = \sum_{i=1}^{k} h_j \frac{f_j}{h_j} = \sum_{j=1}^{k} f_j = n \text{ ou } \sum_{j=1}^{k} h_j f^*_{d_j} = \sum_{j=1}^{k} h_j \frac{f^*_j}{h_j} = \sum_{j=1}^{k} f^*_j = 1.$$

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

Histograma em R

```
> ? co2
> x = as.vector(co2)
> hist(x)
```



Fornece uma ideia sobre a distribuição.

```
> hx = hist(x, right = FALSE, plot = FALSE)
> names(hx)
```

```
[1] "breaks"      "counts"
"intensities" "density"
"mids"        "xname"
"equidist"
```

```
> hx$breaks
```

```
 [1] 310 315 320 325 330 335
340 345 350 355 360 365 370
```

```
> hx$counts
```

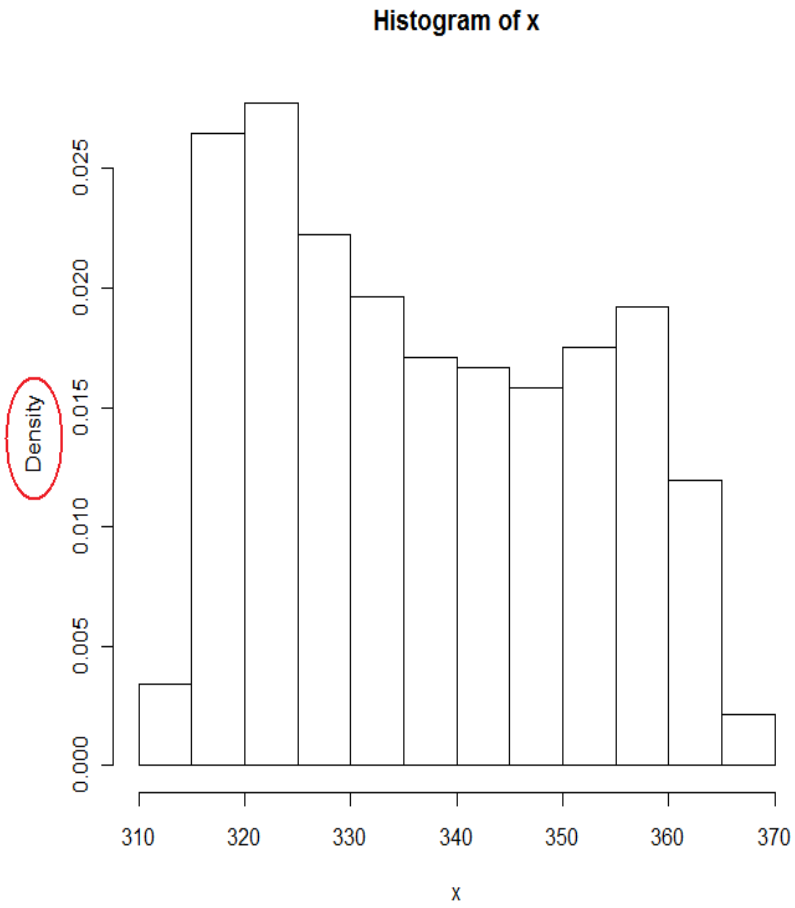
```
 [1]  8 62 65 52 46 40 39 37 41
45 28  5
```

```
> hx$mid
```

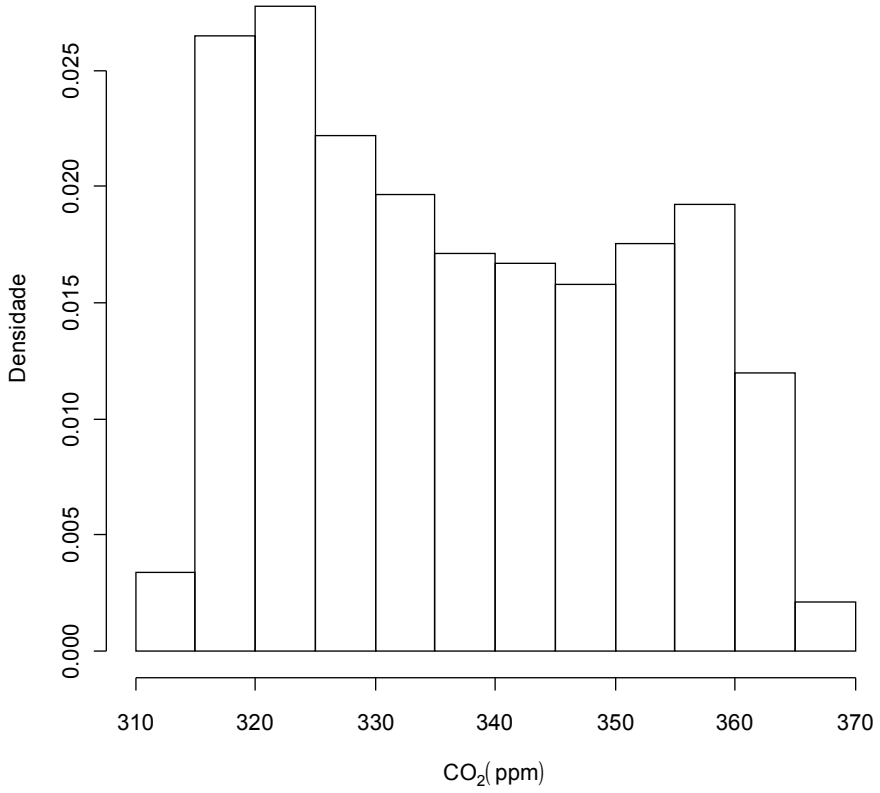
```
 [1] 312.5 317.5 322.5 327.5
332.5 337.5 342.5 347.5 352.5
357.5 362.5 367.5
```

Histograma em R

> hist(x, right = FALSE, freq = FALSE)

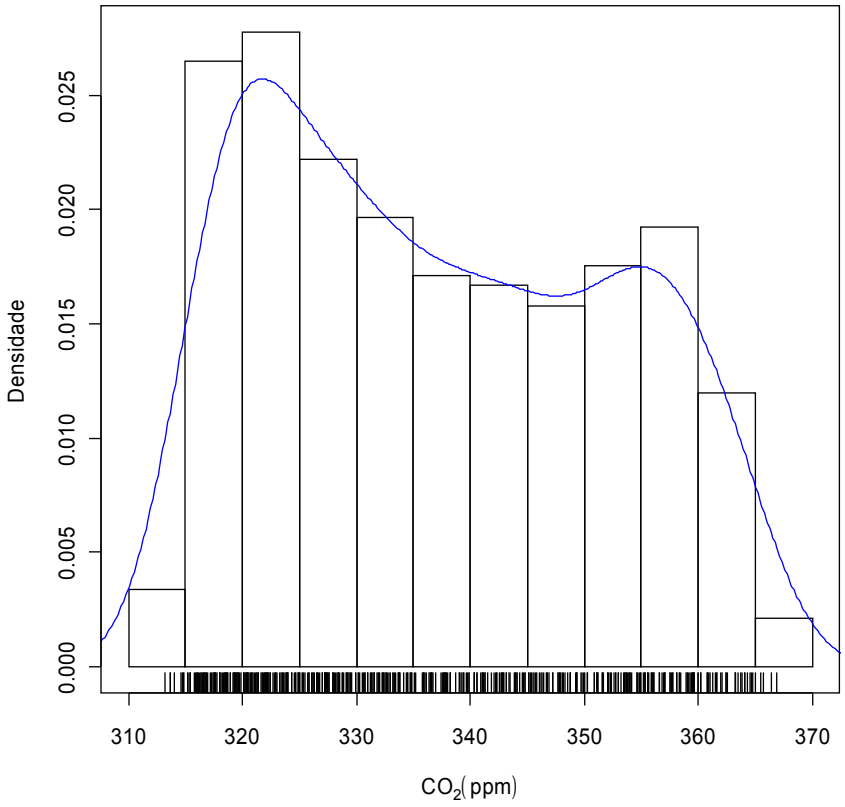


> hist(x, right = FALSE, freq = FALSE, main = "", xlab = expression(CO[2] (ppm)), ylab = "Densidade")



Histograma em R

```
> hist(x, right = FALSE, freq = FALSE, main = "", xlab = expression(CO[2] (ppm)),  ylab = "Densidade")
> rug(x)
> lines(density(x), col = "blue")
> box()
```

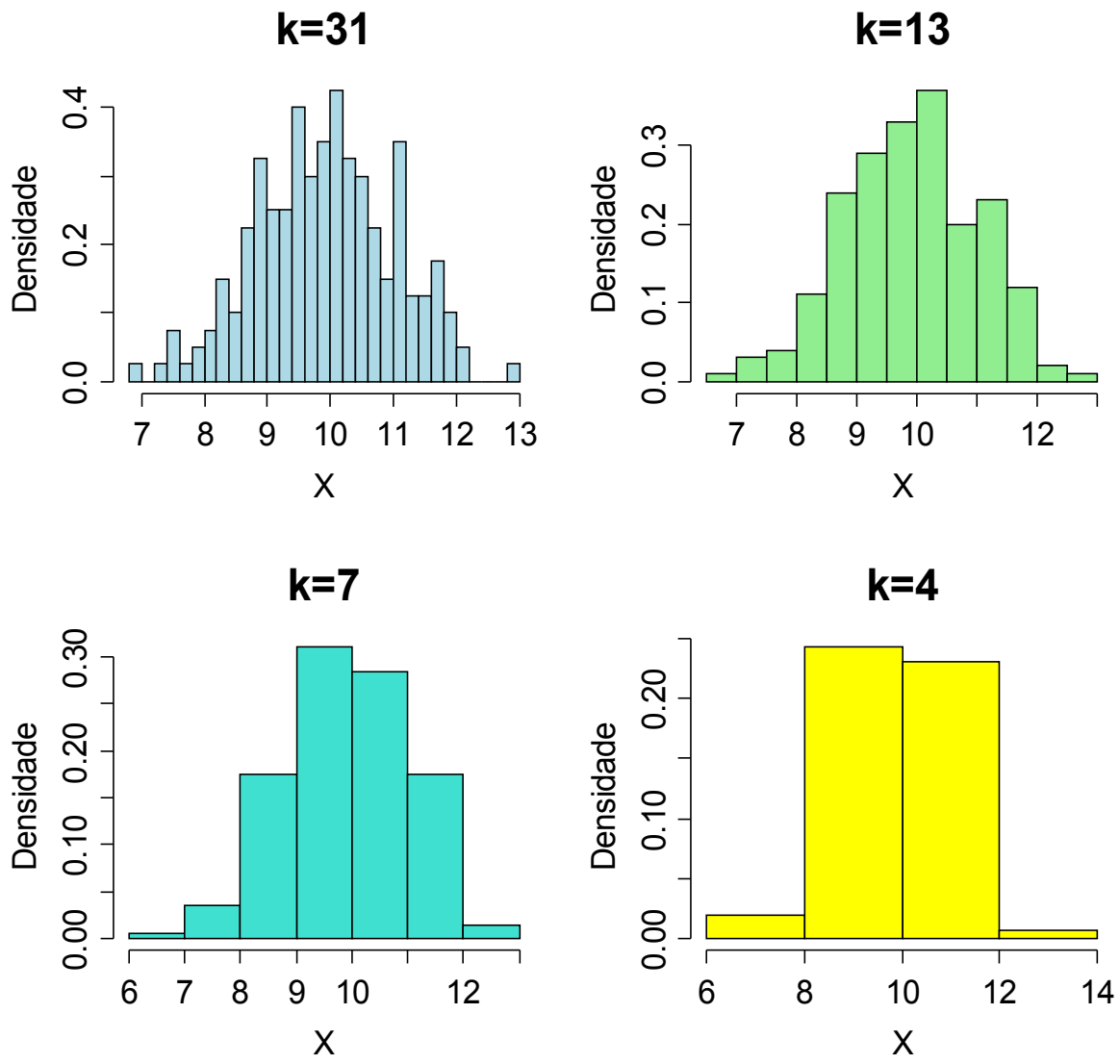


Número de classes: fórmula de Sturges se *breaks* não for especificado.

Outras opções:

1. Fórmula de Scott: *breaks = "Scott"*.
2. Fórmula de Freedman-Diaconis: *breaks = "FD"*.
3. *breaks = número*: nem sempre funciona.
4. *breaks = vetor ordenado* com k + 1 elementos com os limites das classes.

Escolha do número de classes (k)



Histograma humano

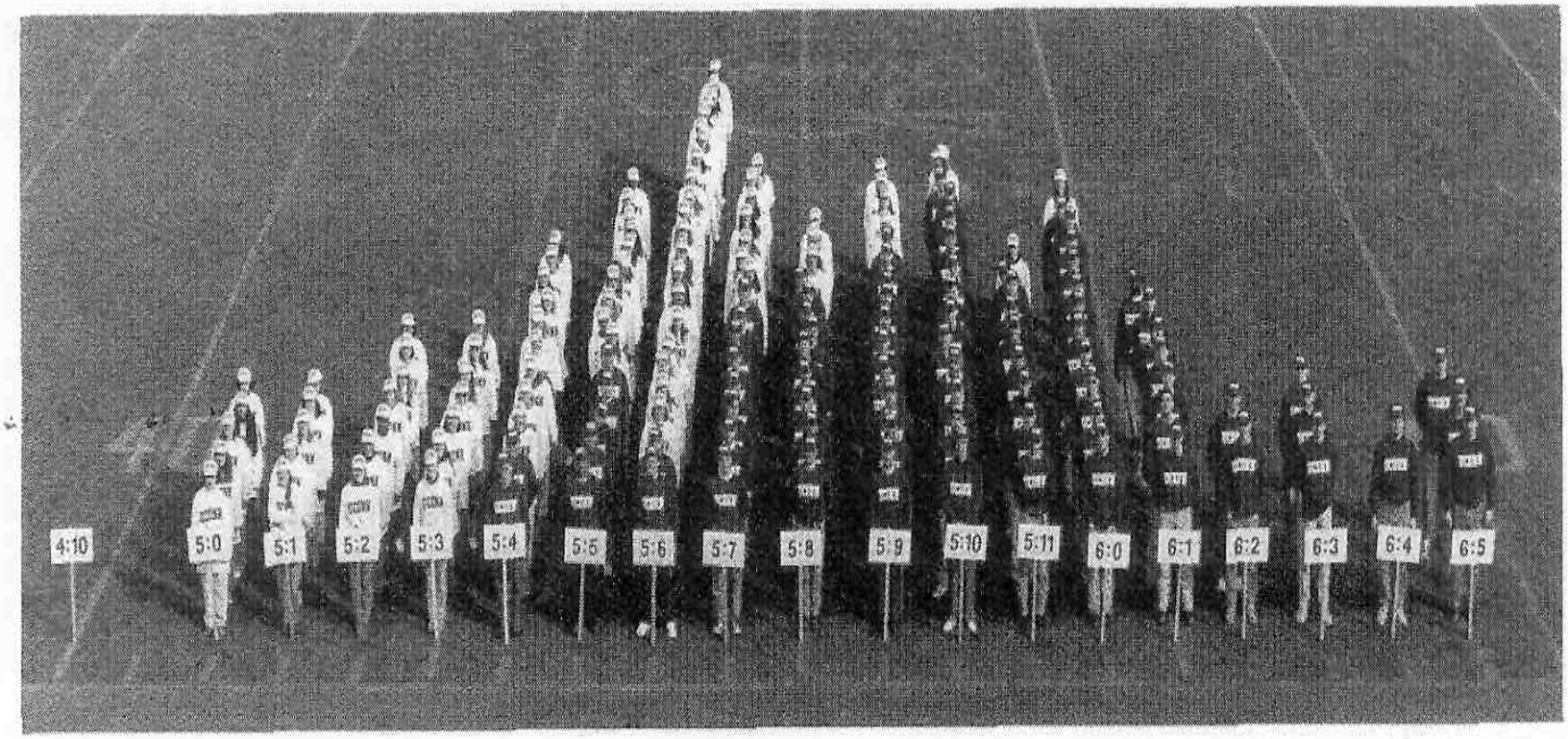


Figure 7. Living histogram of 143 student heights at University of Connecticut.

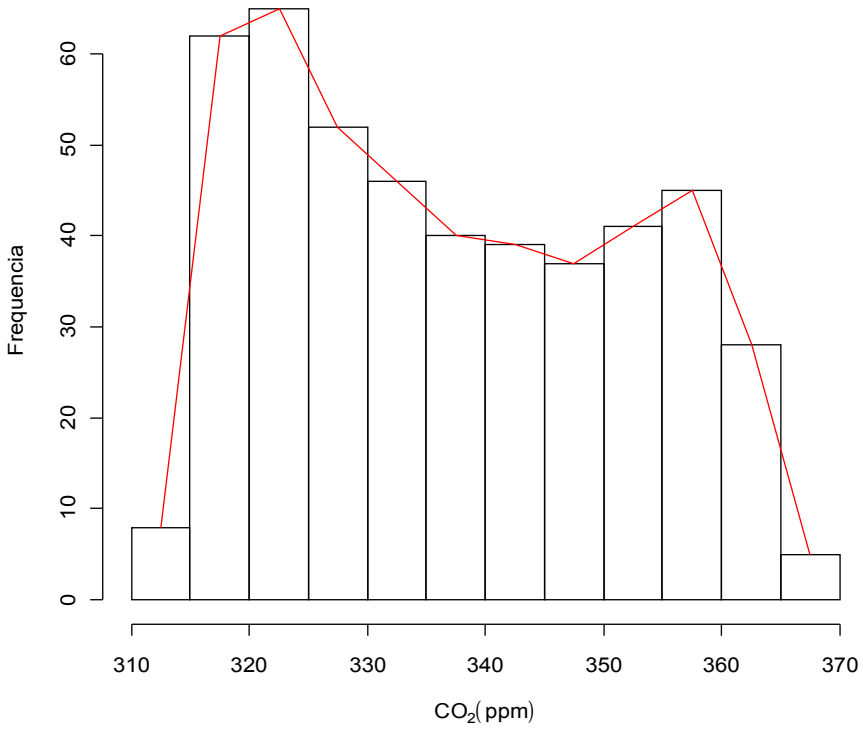
The American Statistican 56(3), 223 – 229, 2002.

Polígono de frequencias

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

```
> hist(x, right = FALSE, main = "", xlab = expression(CO[2] (ppm)), ylab = "Frequencia")
```

```
> lines(hx$mid, hx$counts, col = “red")
```

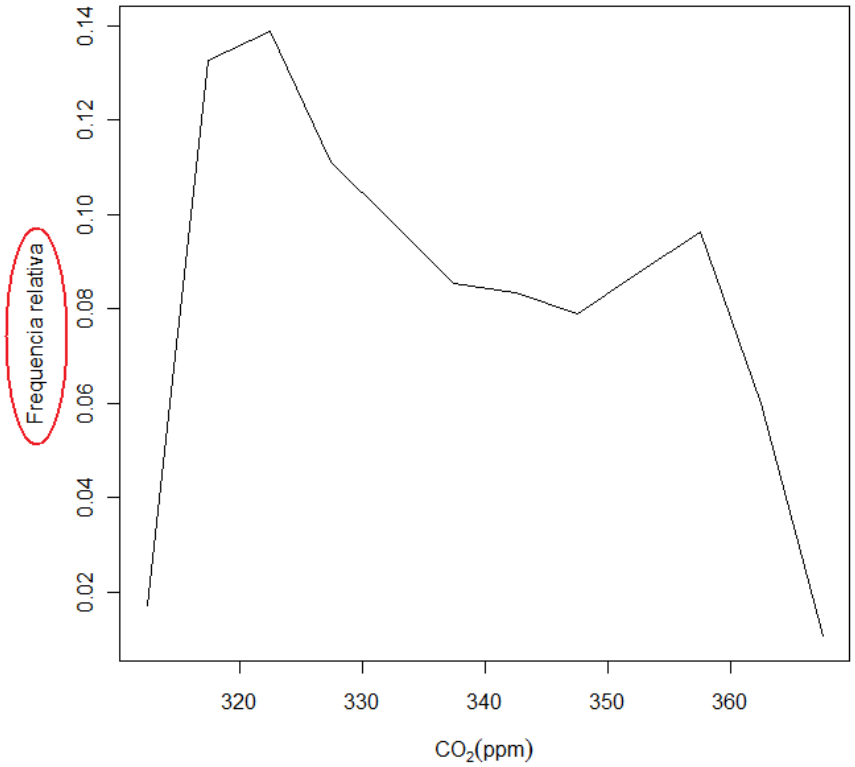



```
> (n = length(x))
```

```
> frel = hx$counts / n
```

```
> plot(hx$mid, frel, type = "l", xlab = expression(CO[2] (ppm)), ylab = “Frequencia relativa")
```

```
[1] 468
```

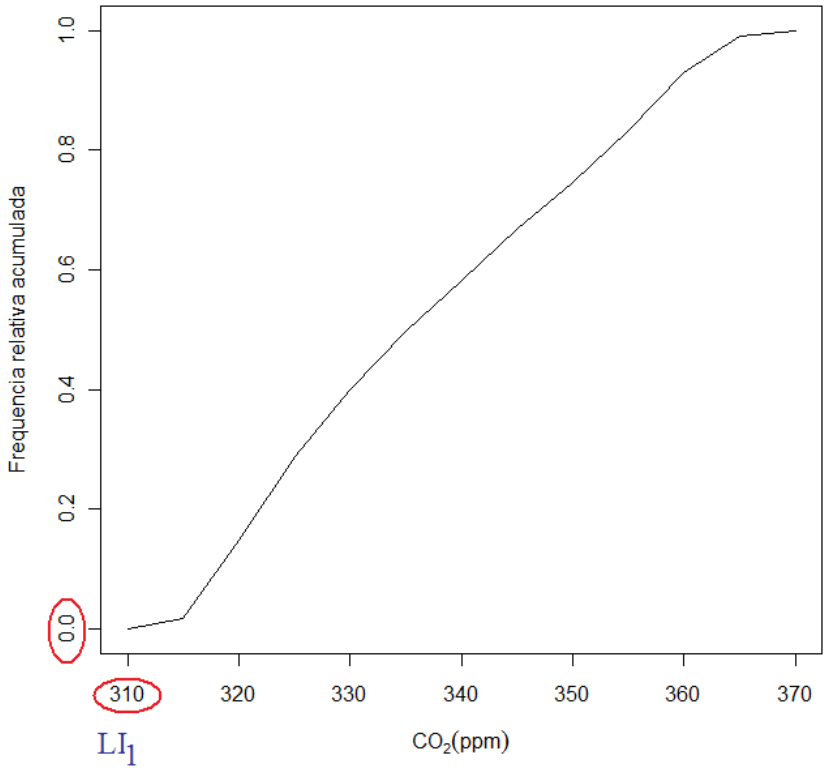


Polígono de frequencias acumuladas (ogiva)

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

```
> Frel = cumsum(frel)

> plot(hx$breaks, c(0, Frel), type = "l", xlab =
expression(CO[2] (ppm)),    ylab =
"Frequencia relativa acumulada")
```




```
> posb = barplot(Frel, col = "white", space= 0,
xlab = expression(CO[2] (ppm)),  ylab =
"Frequencia relativa acumulada")

> lines(posb + posb[1], Frel, col = "red")

> segments(0, 0, (posb[1] + posb[2]) / 2,
Frel[1], col = "red")

> axis(1, c(0, posb + posb[1]), hx$breaks)
```

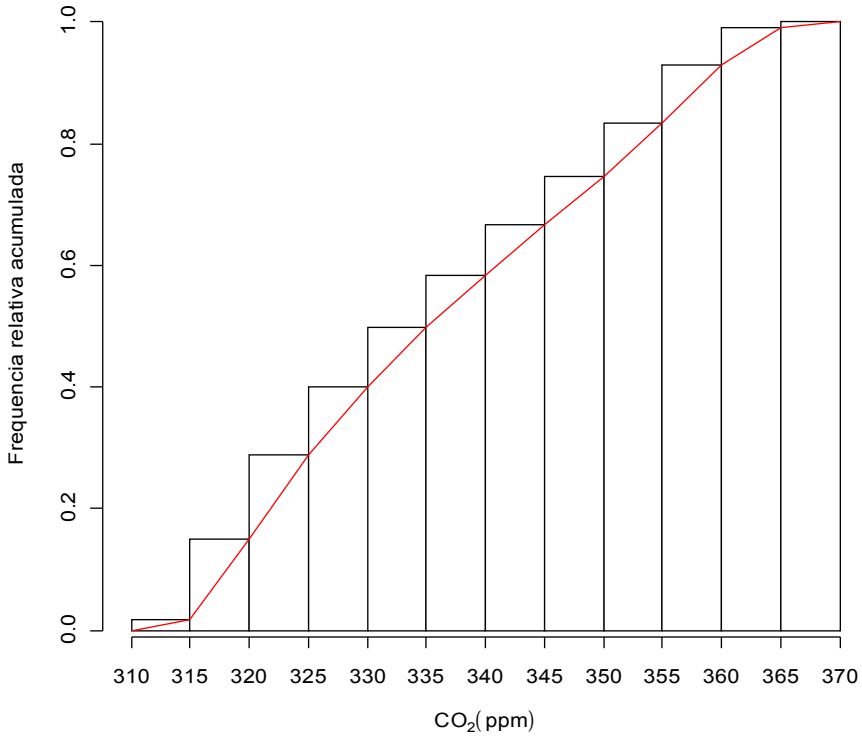


Gráfico de pontos

Cada observação é representada por um ponto.

Não há perda de informação.

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

```
> par(mfrow = c(2, 1))
> stripchart(x, pch = 20, method = "stack", xlab = expression(CO[2] (ppm)),  main = "Sem arredondamento")
> stripchart(round(x, 1), pch = 20, method = "stack", xlab = expression(CO[2] (ppm)),  main = "Com arredondamento")
```

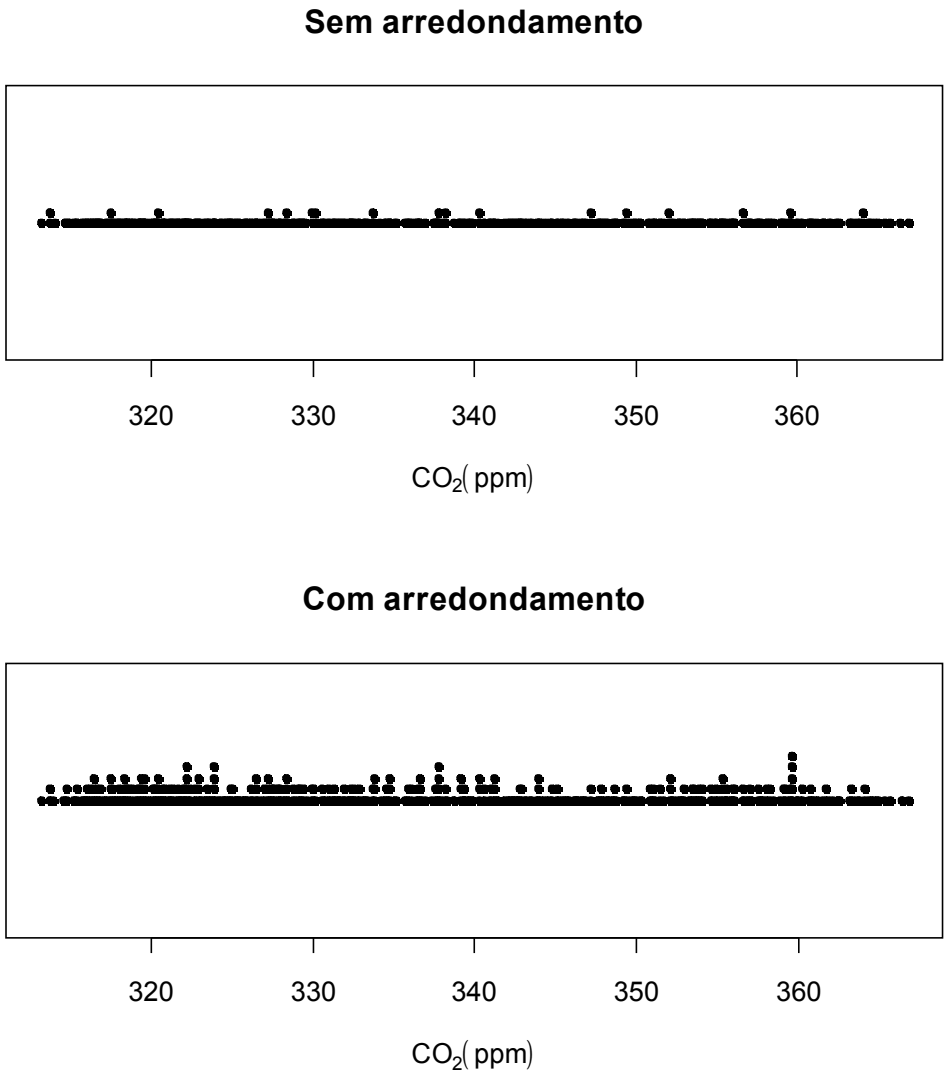


Gráfico de linhas

Utilizado para representar variáveis coletadas com referência a uma unidade de tempo. Chamadas de séries históricas ou séries temporais (*time series*).

Obs. Séries temporais podem ser de variáveis discretas ou qualitativas.

```
> ? AirPassengers
```

```
Monthly Airline Passenger
Numbers 1949-1960
```

```
> class(AirPassengers)
```

```
[1] "ts"
```

```
>  plot(AirPassengers)
```

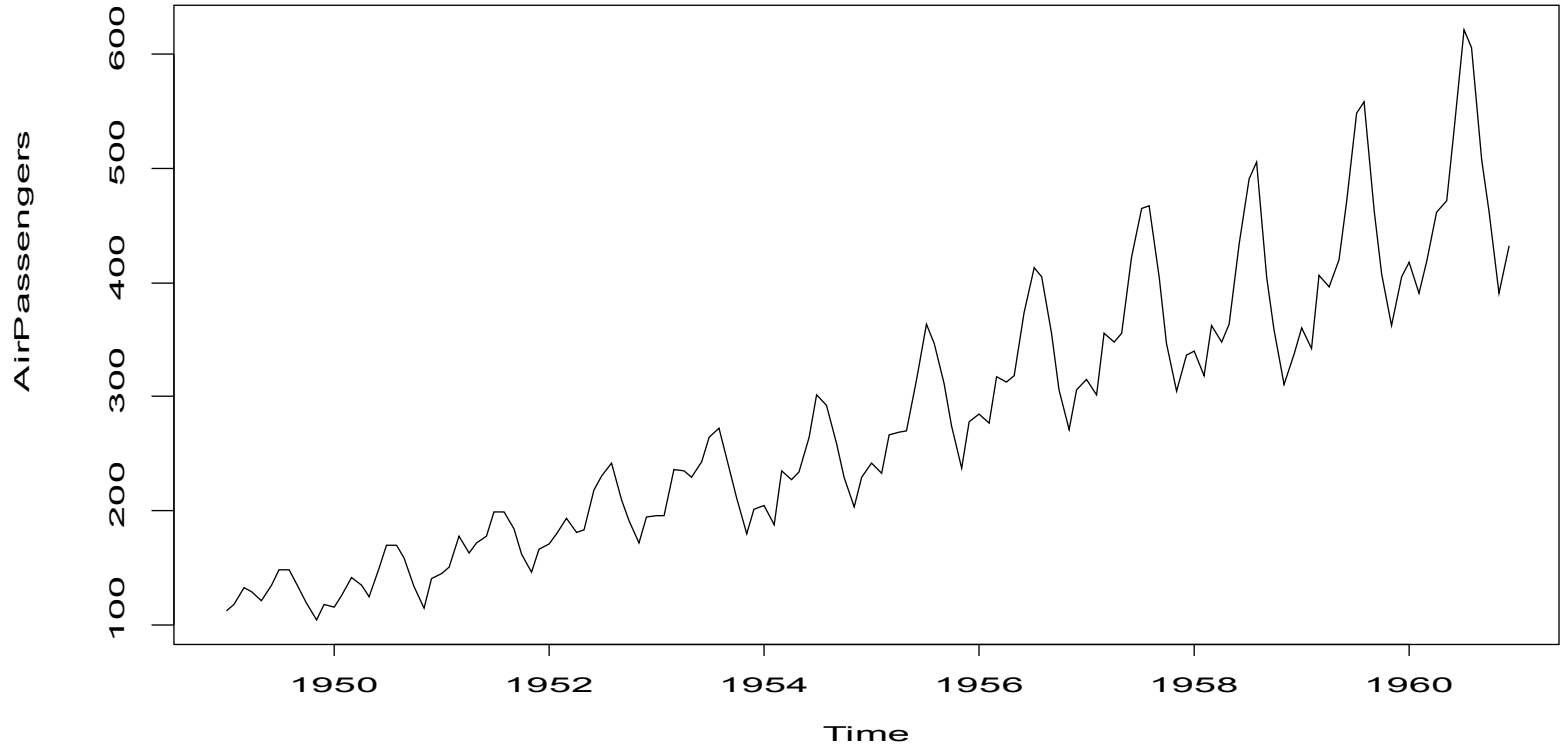


Gráfico de linhas

O Estado de S. Paulo, 28/2/2010.

Gráfico de ramos-e-folhas (*stem-and-leaf plot*)

Representação com nenhuma ou pouca perda de informação.

Cada valor 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 ramos-e-folhas

Notas de 100 alunos em uma certa prova.

> notas

```
5.3  7.0  6.0  7.0  4.4  5.5  9.0  3.1  5.9  4.4  5.5  5.7  3.4  4.8  9.6  7.9
4.7  4.1  7.7  4.2  9.3  3.6  4.6  3.7  8.9  6.0  3.4  7.2  4.2  5.9  5.0  1.8
7.1  5.9  7.3  6.9  3.5  6.4  4.7  4.6  5.2  6.8  8.4  9.3  8.7  4.0  7.6  7.2
3.4  7.8  6.4  4.1  7.9  6.0  5.3  5.3  5.7  5.1  4.0  4.5  8.2  2.6  5.1  5.8
9.0  5.6  5.4  4.1  3.8  5.5  5.6  4.9  8.3  6.8  5.5  5.0  4.6  3.4  6.2  5.1
4.4  6.8 10.0  6.5  7.7  6.1  5.3  6.2  4.6  4.8  8.5  7.2  3.5  2.5  5.3  6.5
4.6  3.9  6.6  7.7
```

> sort(notas)

```
1.8  2.5  2.6  3.1  3.4  3.4  3.4  3.4  3.5  3.5  3.6  3.7  3.8  3.9  4.0  4.0
4.1  4.1  4.1  4.2  4.2  4.4  4.4  4.4  4.5  4.6  4.6  4.6  4.6  4.6  4.7  4.7
4.8  4.8  4.9  5.0  5.0  5.1  5.1  5.1  5.2  5.3  5.3  5.3  5.3  5.3  5.4  5.5
5.5  5.5  5.5  5.6  5.6  5.7  5.7  5.8  5.9  5.9  5.9  6.0  6.0  6.0  6.1  6.2
6.2  6.4  6.4  6.5  6.5  6.6  6.8  6.8  6.8  6.9  7.0  7.0  7.1  7.2  7.2  7.2
7.3  7.6  7.7  7.7  7.7  7.8  7.9  7.9  8.2  8.3  8.4  8.5  8.7  8.9  9.0  9.0
9.3  9.3  9.6 10.0
```

Parte fracionária: folhas. Parte inteira: ramos.

Gráfico de ramos-e-folhas

> stem(notas)

```
  The decimal point is at the |

   1 | 8
   2 | 56
   3 | 14444556789
   4 | 00111224445666677889
   5 | 00111233333455566778999
   6 | 000122445568889
   7 | 00122236777899
   8 | 234579
   9 | 00336
  10 | 0
```

Fornece uma ideia sobre a distribuição.

> stem(notas, density = 2)

```
  The decimal point is at the |

   1 | 8
   2 |
   2 | 56
   3 | 14444
   3 | 556789
   4 | 0011122444
   4 | 56666677889
   5 | 001112333334
   5 | 555566778999
   6 | 00012244
   6 | 5568889
   7 | 0012223
   7 | 6777899
   8 | 234
   8 | 579
   9 | 0033
   9 | 6
  10 | 0
```