1. ANÁLISE EXPLORATÓRIA E ESTATÍSTICA DESCRITIVA

2019

Estatística Descritiva e Análise Exploratória

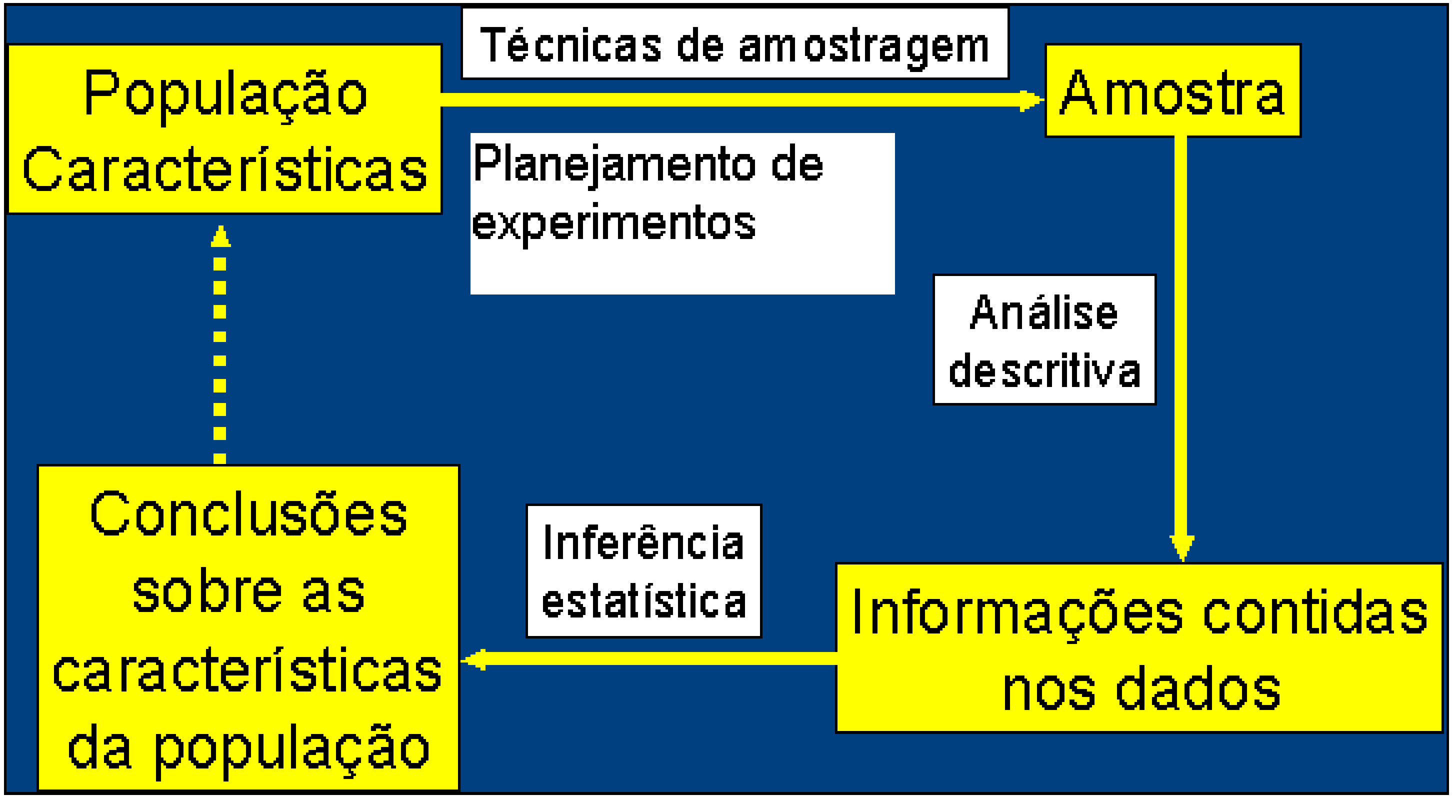
Etapas iniciais. Utilizadas para descrever e resumir os dados.

A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou estas áreas da Estatística.

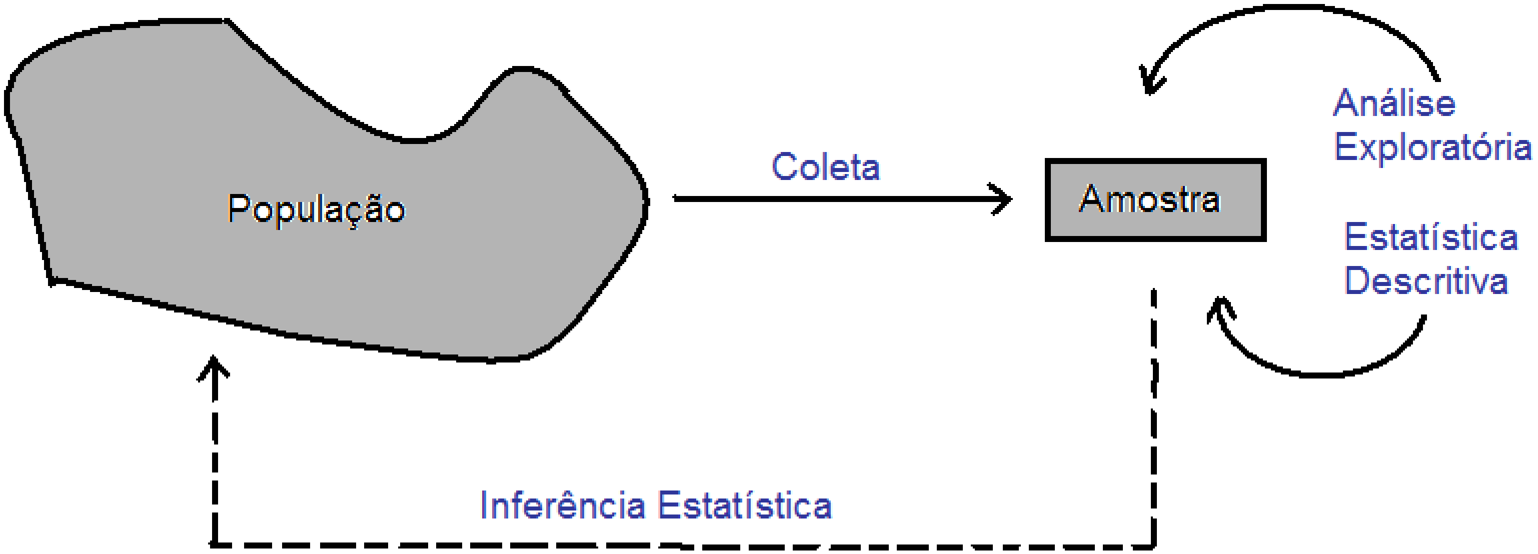
Probabilidade

Permite estudar os fenômenos aleatórios, ou seja, aqueles em que está presente a incerteza sobre os seus resultados.

Estatística



Estatística



O que é Estatística ?

Para muitos, Estatística não passa de conjuntos de tabelas de dados numéricos. Os estatísticos são pessoas que coletam esses dados.

A Estatística originou-se com a coleta de dados e a construção de tabelas para os governos.
A situação evoluiu e esta coleta de dados representa somente um dos aspectos da Estatística.

Definição de Estatística

A Estatística é uma ciência baseada na **Teoria da Probabilidade**, cujo objetivo principal é nos auxiliar a tomar decisões ou tirar conclusões em situações de incerteza, a partir de dados.

População: conjunto de todas as unidades que são de interesse em um certo estudo.

Amostra: qualquer subconjunto da população selecionado de acordo com certas regras.

Censo: estudo que inclui todos os elementos da população.

Coleta

Experimento planejado

Efeito de um ou mais fatores sobre outro(s).

Interferência do pesquisador.

Controle sobre fatores externos.

Levantamento observacional

Dados são coletados “como estão”.

Não há interferência do pesquisador.

Levantamento amostral (*survey*)

População bem definida.

Protocolo de coleta.

Amostragem

Uma área importante em muitas aplicações estatísticas é a da **Tecnologia de Amostragem.**

Exemplos:

- Pesquisa de mercado
- Pesquisa de intenção de voto (pesquisa eleitoral)
- Avaliação do impacto de uma obra junto à população

O que fazer com os dados coletados?

1ª etapa: Estatística Descritiva e Análise Exploratória

Medidas resumo, tabelas e gráficos.

Obs. Se *x* representa uma variável, uma amostra com valores $x_1, x_2, ..., x_n$ é chamada de conjunto de dados.

n é o tamanho da amostra.

Variável

Qualquer característica de interesse associada aos elementos de uma população.

Classificação de variáveis

Qualitativa	Nominal	Cor, tipo de máquina
	Ordinal	Classe social, grau de desgaste
Quantitativa	Discreta	Número de acidentes, número de defeitos em um item
	Contínua	Peso, viscosidade, pressão

Exemplo. Estudo de resistência.

Observação	Espessura	Tipo de cola	Resistência
1	13	1	46,5
2	14	1	45,9
3	12	1	49,8
4	12	1	46,1
5	14	1	44,3
6	12	2	48,7
7	10	2	49,0
8	11	2	50,1
9	12	2	48,5
10	14	2	45,2
11	15	3	46,3
12	14	3	47,1
13	11	3	48,9
14	11	3	48,2
15	10	3	50,3
16	16	4	44,7
17	15	4	43,0
18	10	4	51,0
19	12	4	48,1
20	11	4	48,6

Fonte: Montgomery, D. C. (2005), Design and Analysis of Experiments, 6th Edition, Wiley: New York

Medidas resumo

Medidas de posição: moda, média, mediana, percentis, quartis.
(medidas de tendência central: três primeiras)

Medidas de dispersão: amplitude, intervalo interquartil, variância, desvio padrão, coeficiente de variação.

Moda (Mo): É o valor (ou atributo) que ocorre com maior frequência.

Ex. Dados: 4,5,4,6,5,8,4,4
mo = 4

Obs. 1. Nem sempre a moda existe.
2. Pode haver mais de uma moda.

Média: $$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ex. Dados: 2,5,3,7,11

$\bar{X}$ = (2+5+3+7+11)/5 = 5,6

Mediana (Md)

A mediana é o valor que ocupa a posição central de um conjunto de n valores ordenados.
Posição da mediana: pm = (n+1)/2

Ex. Dados: 2, 26, 3, 7, 8 (n = 5)

Dados ordenados: 2,3,7,8, 26 => pm = (5+1)/2=3
=> Md = 7

Ex. Dados: 2, 15, 2, 1, 8, 7 (n = 6)

Dados ordenados: 1, 2, 2, 7, 8, 15 => pm = (6+1)/2=3,5
=> Md = (2+7) / 2 = 4,5 (média dos elementos nas posições 3 e 4).

Quantis (*quantiles*)

O quantil de ordem p ($0 < p < 1$), em um conjunto de dados com n observações, é o valor que ocupa a posição p x (n+1) nos dados ordenados.

px100% das observações são ≤ quantil de ordem p e

(1 - p)x100% das observações são ≥ quantil de ordem p.

Casos particulares:

Quantil 0,5 = mediana ou segundo quartil (md)

Quantil 0,25 = primeiro quartil (Q1)

Quantil 0,75 = terceiro quartil (Q3)

Exemplos

Ex. 1. 1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7 (n = 10)

Posição da Md: $0{,}5 \times (n+1) = 0{,}5 \text{x} 11 \Rightarrow Md = (3+3{,}1)/2 = 3{,}05$

Posição de Q1: $0{,}25 \times 11 = 2{,}75 \Rightarrow Q1 = (2+2{,}1)/2 = 2{,}05$

Posição de Q3: $0{,}75 \times 11 = 8{,}25 \Rightarrow Q3 = (3{,}7+6{,}1)/2 = 4{,}9$

Ex. 2. 0,9 1,0 1,7 2,9 3,1 5,3 5,5 12,2 12,9 14,0 33,6 (n = 11)

Md = 5,3

Q1 = 1,7

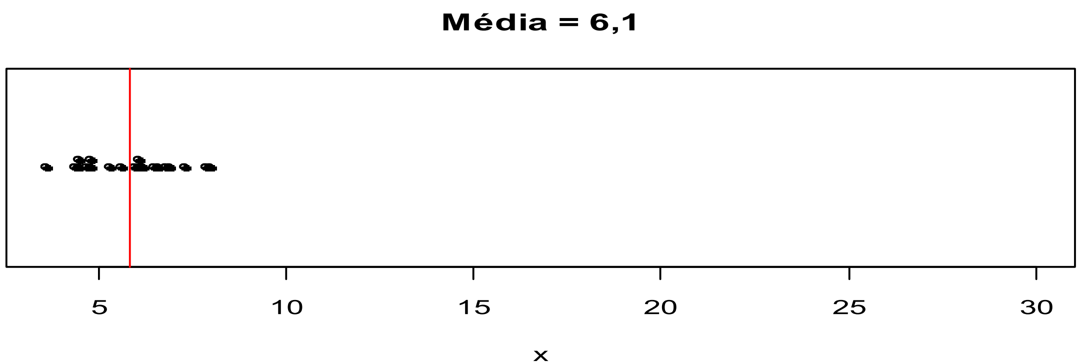
Q3 = 12,9

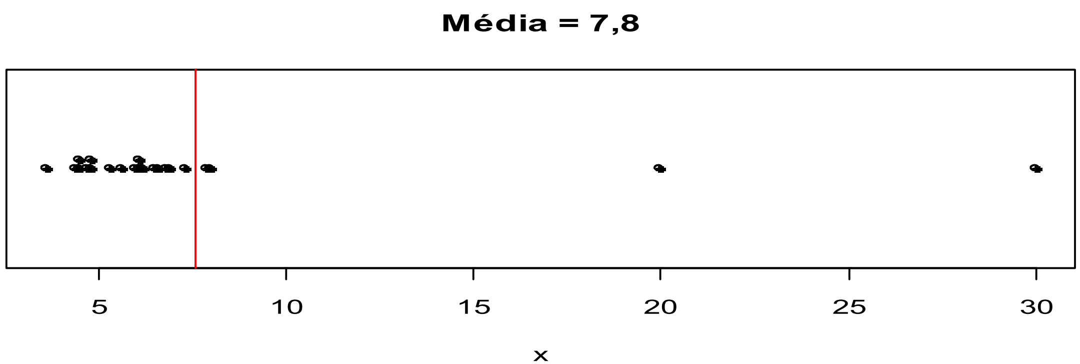
Moda, mediana e média (*mode, median and mean*)

A moda não é muito utilizada com variáveis quantitativas.

Se a variável for qualitativa nominal, a moda é a única medida de posição.

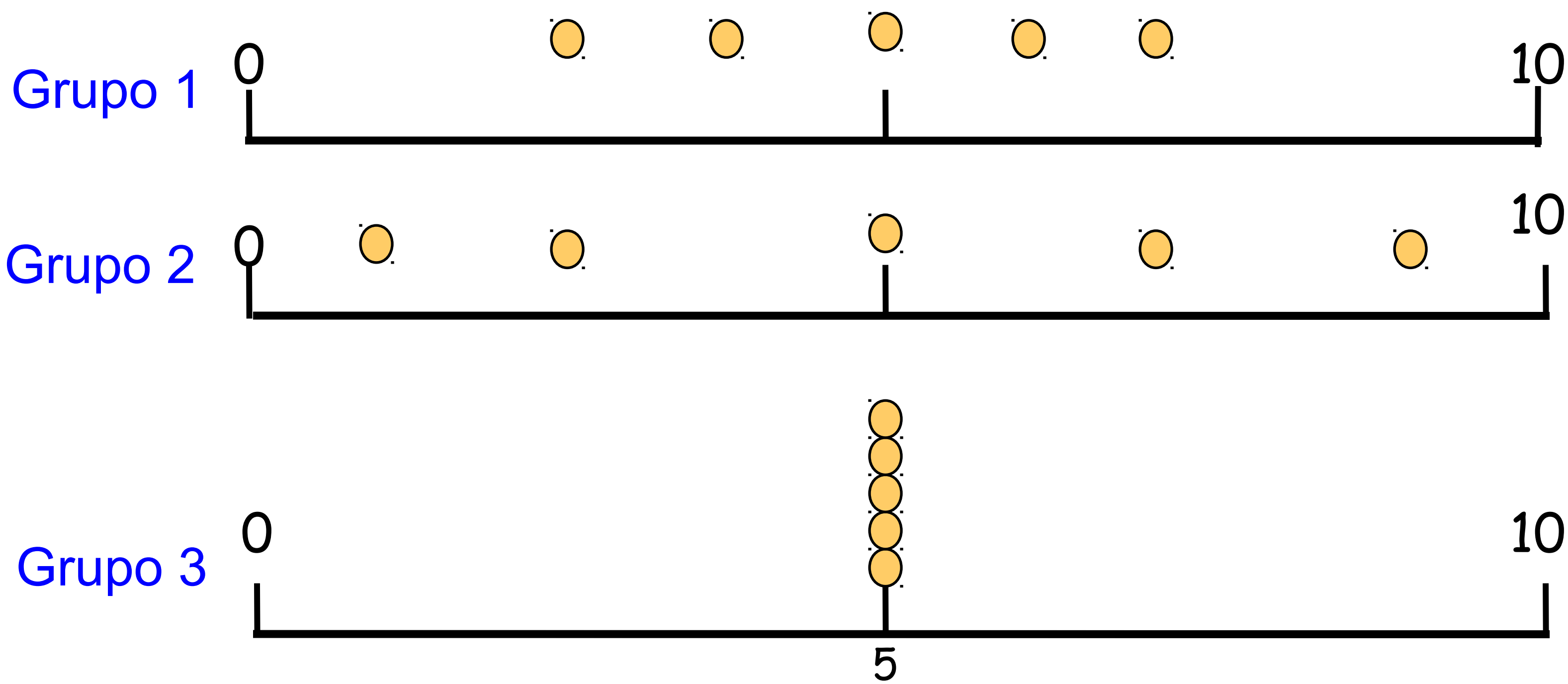
A mediana é mais resistente do que a média. É menos afetada pela presença de valores extremos.





Obs. Os quantis também são chamados de separatrizes.

Considere as notas de uma prova aplicada a três grupos de alunos:
Grupo 1: 3, 4, 5, 6, 7; Grupo 2: 1, 3, 5, 7,9; e Grupo 3: 5,5,5,5,5.



$$\bar{x}_1 = \bar{x}_3 = \bar{x}_3 = 5;\ Md_1 = Md_3 = Md_3 = 5$$

Medidas de dispersão

Finalidade: encontrar um valor que resuma a variabilidade de um conjunto de dados.

Amplitude (A): A = MAX - min

Para os grupos anteriores (lâmina 18), temos

Grupo 1: A = 4

Grupo 2: A = 8

Grupo 3: A = 0

Intervalo ou amplitude interquartil (d_q) (*interquartile range*)

É a diferença entre o terceiro quartil e o primeiro quartil:

$$d_q = Q_3 – Q_1.$$

Ex. 1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7

$Q_1 = 2{,}05$ e $Q_3 = 4{,}9$.

$d_q = Q_3 – Q_1 = 4{,}9 – 2{,}05 = 2{,}85$.

Obs. d_q é uma medida mais resistente do que A.

Variância (s^2) (*variance*)

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Desvio padrão (s) (*standard deviation*)

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

Obs. O desvio padrão tem a mesma unidade da variável *x*.

Cálculo da variância para o grupo 1 (lâmina 18):

Grupo 1: 3, 4, 5, 6, 7: Vimos que $\bar{x}=5$

$$S^2=\frac{(3-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-5)^2}{5-1}=\frac{10}{4}=2{,}5$$

Desvio padrão:

$$\text{Grupo 1: } s^2=2{,}5 \Rightarrow s=1{,}58$$

$$\text{Grupo 2: } s^2=10 \Rightarrow s=3{,}16$$

$$\text{Grupo 3: } s^2=0 \Rightarrow s=0$$

Propriedades:

$x_1, \ldots, x_n$ uma amostra com média $\bar{x}$ e variância s_x^2.

Transformação (posição e escala): $y_i = a + b\,x_i$, $i = 1,\ldots,n$.

$$\bar{y} = a + b\,\bar{x},$$

$$s_y^2 = b^2 s_x^2 \quad \text{e} \quad s_y = |b|\,s_x.$$

Coeficiente de variação (CV)

É uma medida de dispersão relativa.

Exprime a variabilidade em relação à média.

$$CV = \frac{S}{|\bar{x}|} \times 100$$

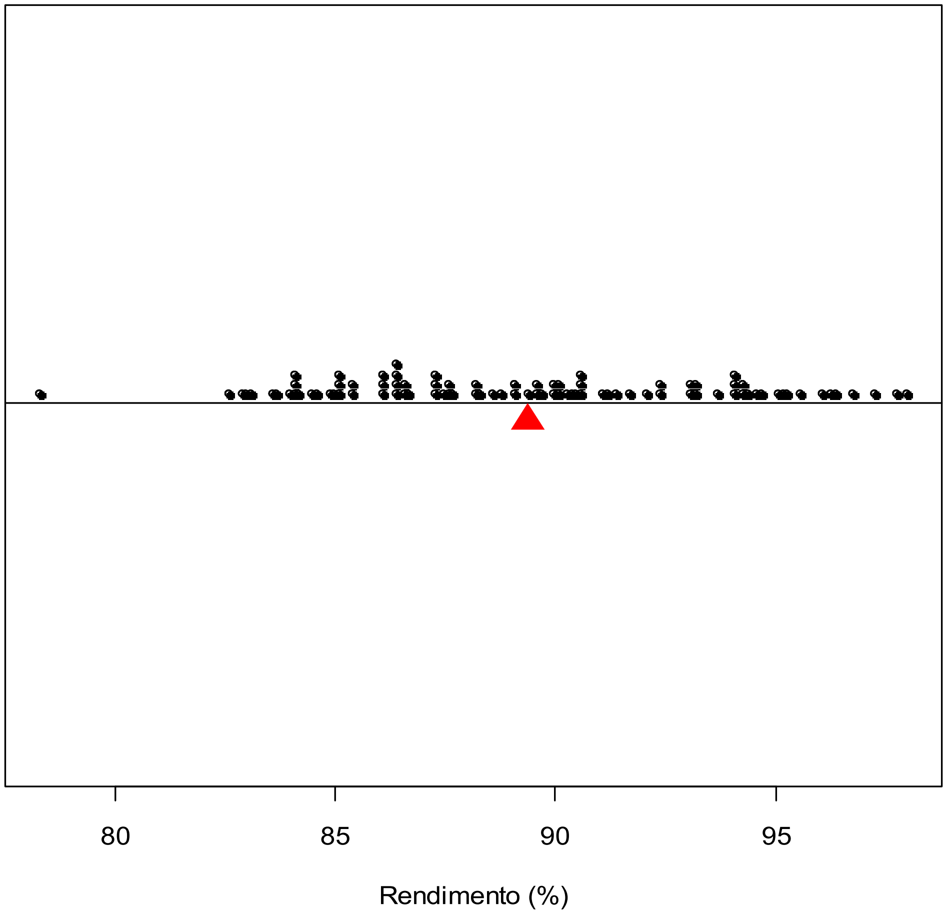
se $\bar{x} \neq 0$.

Exemplo. Altura e peso de 15 peças metálicas cilíndricas

	Média	Desvio padrão	Coeficiente de variação
Altura	1,143 m	0,063 m	5,5%
Peso	50 kg	6 kg	12%

Conclusão. O peso das peças métalicas apresenta variabilidade relativa aproximadamente duas vezes maior do que a altura.

Exemplo com gráfico de pontos (n = 90)



Cada observação é representada por um ponto.

Havendo repetições, os pontos são empilhados.

Propriedade:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

Uma das formas de organizar e resumir a informação contida em dados observados é por meio de tabelas de frequências e gráficos.

A frequência de um valor da variável é o número de vezes que este valor ocorre no conjunto de dados.

Tabela de frequências. Tabela com os diferentes valores de uma variável (ou intervalos de valores) e suas respectivas frequencias.

1. Variáveis qualitativas. Tabela de frequências dos diferentes valores da variável.
Representação gráfica: gráfico de barras, de Pareto e gráfico de setores (“de pizza”).

Exemplo. Variável “Tipo de cimento utilizado” (qualitativa nominal)

	Tipo de cimento	Contagem	f_i	f_{r_i}
	C1		12	0,3333
	C2		18	0,5000
	C3		6	0,1667
	Total		n = 36	1,0000

f_i : frequência absoluta do valor i (número de observações com tipo de cimento = i), i $\in$ {C1, C2, C3}.

$f_{r_i} = \dfrac{f_i}{n}$: frequência relativa do valor i.

Elementos de um gráfico

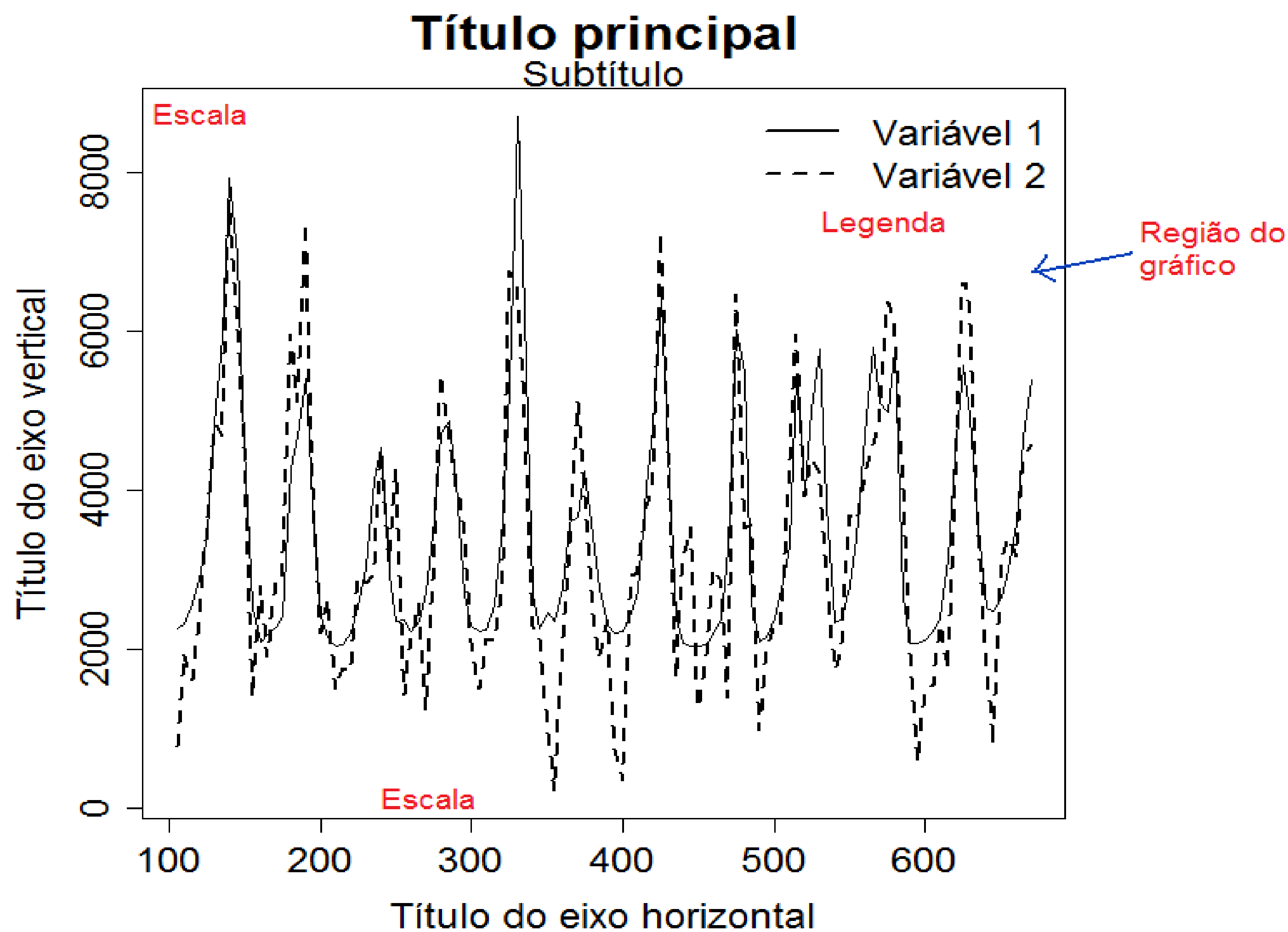
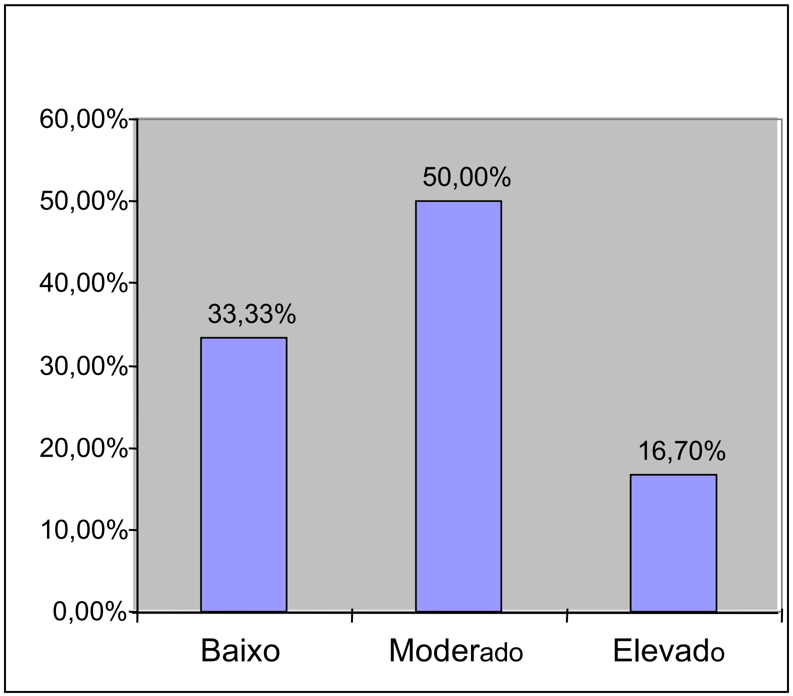


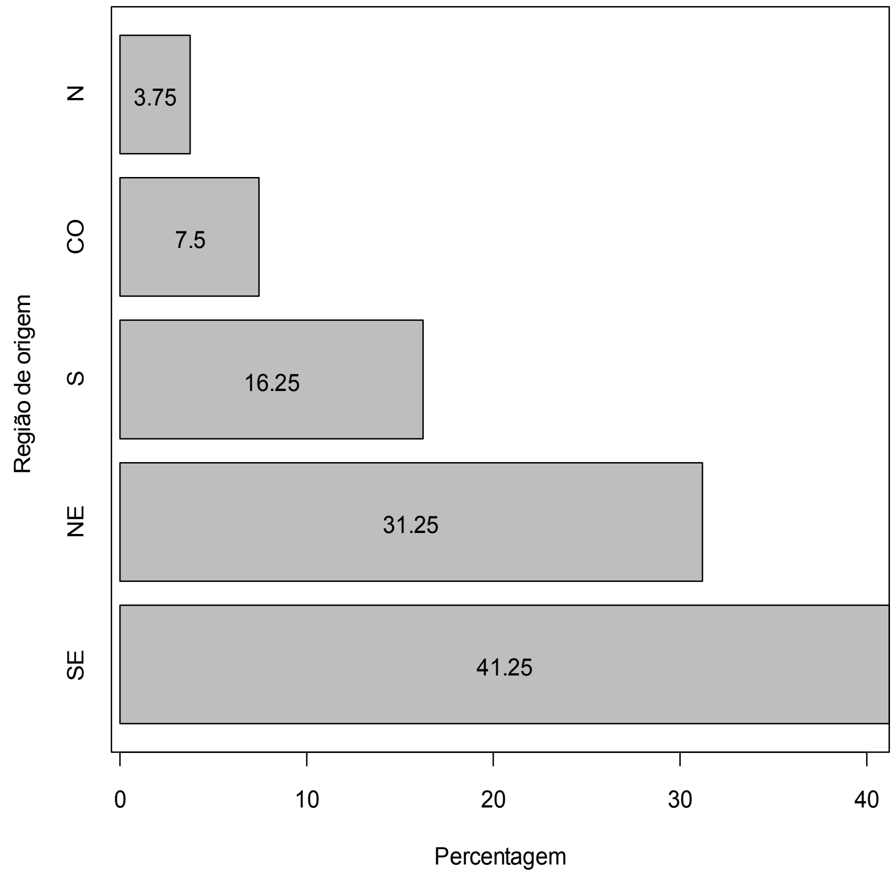
Figura 1. Descrição do gráfico.

Representação gráfica de variáveis qualitativas

Gráfico de barras: retângulos verticais (ou horizontais) espaçados com alturas (ou bases) iguais às frequencias dos valores da variável.



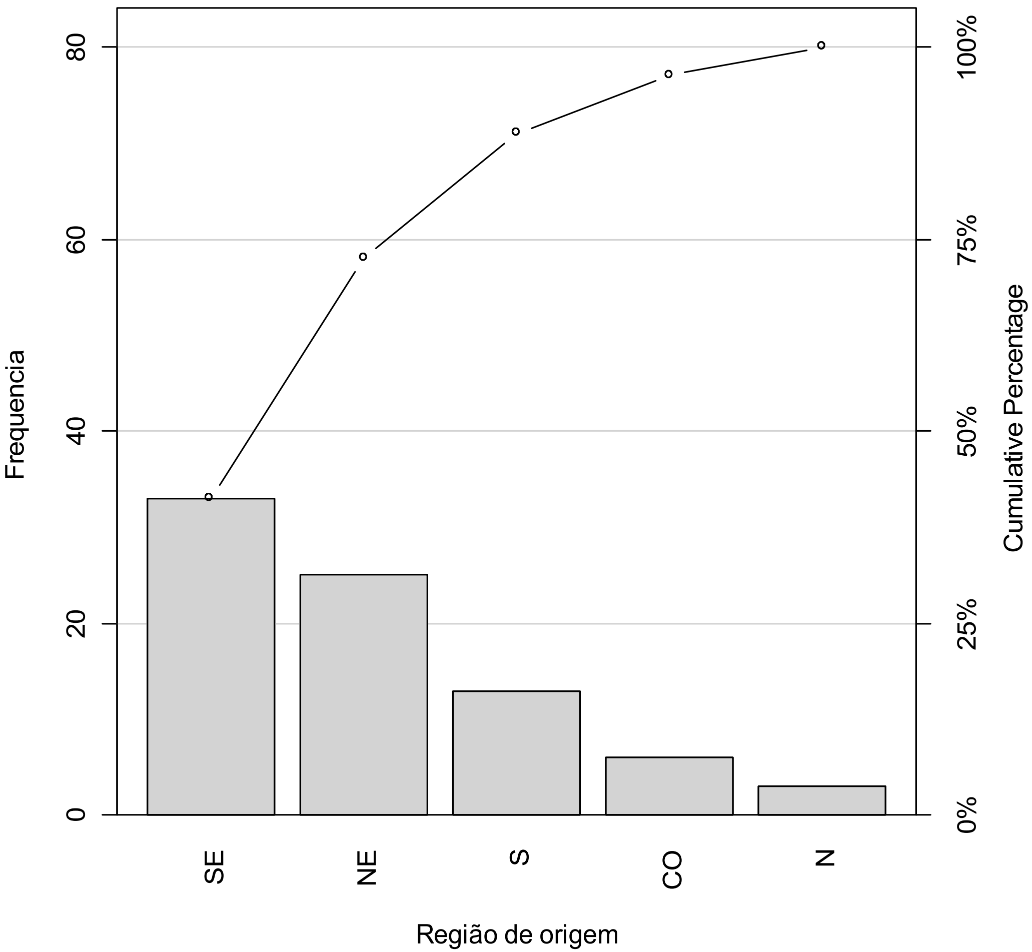
Grau de deterioração



Obs. Os valores no eixo horizontal estão ordenados.

Gráfico de Pareto

Gráfico de barras com os valores da variável em ordem decrescente de frequencias e com as frequencias relativas acumuladas no segundo eixo vertical.



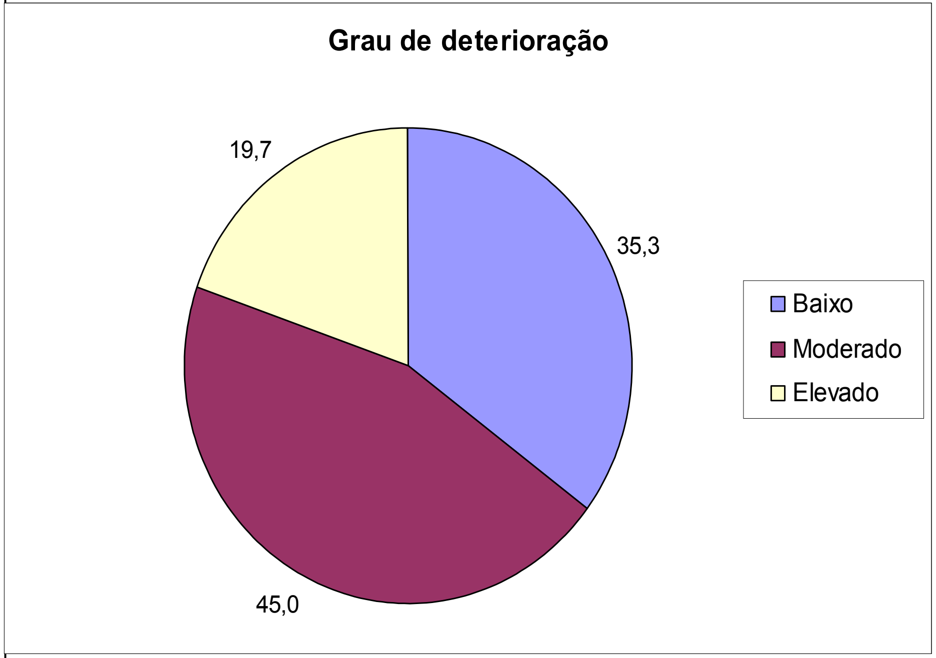
Frequência relativa acumulada do valor x_i:

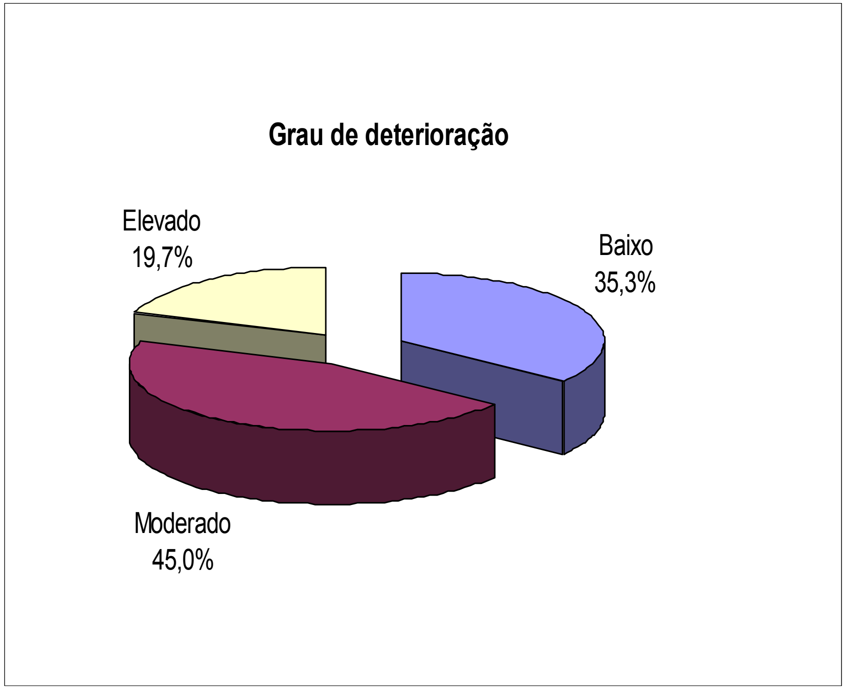
$$F_{ri} = f_{r1} + f_{r2} + \cdots + f_{ri} = \sum_{j=1}^{i} f_{rj}.$$

Gráficos de setores ("de pizza")

Gráfico circular utilizado para destacar a composição das partes de um todo.

O ângulo central de cada setor é proporcional à frequencia representada (usualmente em %).





2.1 Discretas. Organizam-se mediante tabelas de frequências e a representação gráfica é mediante gráfico de pontos, de barras ou de linha.

Frequência relativa do valor x_i : $f_{ri} = f_i / n$.

Frequência acumulada do valor x_i:

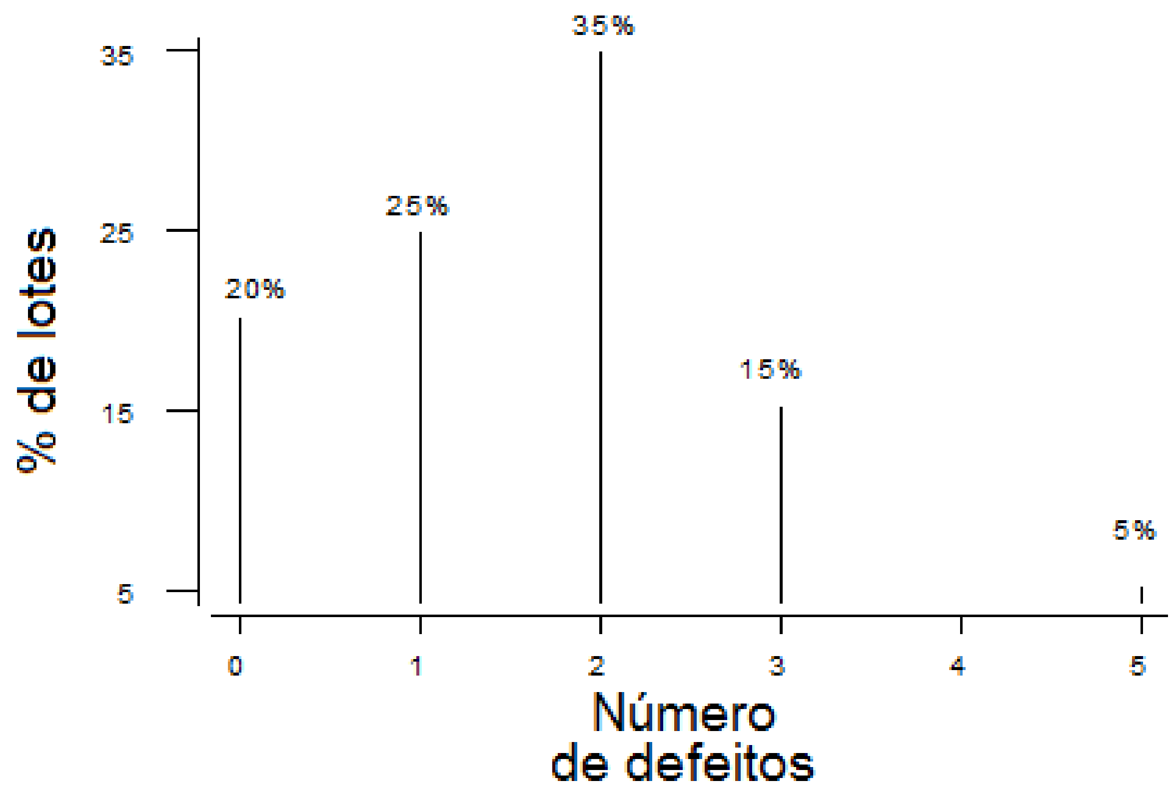
$$F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j$$

Exemplo. Número de defeitos em lotes de produtos.

Distribuição de frequências do número de defeitos por lote.

i	Número de defeitos (X_i)	Número de lotes (f_i)	% de lotes (f_{ri})
1	0	4	20%
2	1	5	25%
3	2	7	35%
4	3	3	15%
5	5	1	5%
Total		20	100%

Representação gráfica



Medidas de posição e dispersão para variáveis quantitativas discretas agrupados em tabela de frequências:

Média:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{n} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

Exemplo. Determine o número médio de defeitos por lote.

$$\bar{x} = \frac{0 \times 4 + 1 \times 5 + 2 \times 7 + 3 \times 3 + 5 \times 1}{20} = \frac{33}{20} = 1{,}65$$

Mediana:

n = 20: pm = (20+1) / 2 = 10,5 =>

Md = média dos valores com frequencias acumuladas iguas a 10 e 11

= (2 + 2) / 2 = 2 (lâmina 33).

Moda = ?

Variância:

$$s^2 = \frac{(x_1-\overline{x})^2 f_1 + (x_2-\overline{x})^2 f_2 + \cdots + (x_k-\overline{x})^2 f_k}{n-1} = \frac{\sum_{i=1}^{k} (x_i-\overline{x})^2 f_i}{n-1}$$

Exemplo.

$$s^2 = \frac{4(0-1{,}65)^2 + 5(1-1{,}65)^2 + 7(2-1{,}65)^2 + 3(3-1{,}65)^2 + (5-1{,}65)^2}{19}$$

$$\frac{16{,}3125}{19} = 0{,}859$$

Desvio padrão: $s = \sqrt{s^2} = 0{,}927$

Coeficiente de variação: $CV = \frac{s}{|\overline{x}|} \times 100\% = \frac{0{,}92}{1{,}65} \times 100\% = 55{,}8\%$

2.2 Construção de tabelas de frequências para variáveis contínuas

- Escolha o número de intervalos ou classes (k)
- Identifique o menor valor (min) e o valor máximo (MAX) dos dados.
- Calcule a amplitude (A): A = MAX – min.
- Calcule a amplitude de classe (h): h = A / k.
- Obtenha os limites inferior (LI) e superior (LS) de cada intervalo.

$1^{\underline{o}}$ intervalo:

Limite inferior: $\mathrm{LI}_1 = \min$

Limite superior: $\mathrm{LS}_1 = LI_1 + h$

$2^{\underline{o}}$ intervalo:

Limite inferior: $\mathrm{LI}_2 = LS_1$

Limite superior: $\mathrm{LS}_2 = LI_2 + h$

...

i-ésimo intervalo:

Limite inferior: $\mathrm{LI}_i = LS_{i-1}$

Limite superior: $\mathrm{LS}_i = LI_i + h$

Prossiga até que seja obtido um intervalo que contenha o valor máximo (MAX).

Obs. Muitas vezes, por conveniência, arrredondamos os valores de h e/ ou LI_1.

Tabela de de frequências com as colunas:

- Número de ordem de cada intervalo (i)
- Limites de cada intervalo. Os intervalos são fechados à esquerda e abertos à direita. Notação: $\vdash$

Ponto médio (ou marca de classe) de cada intervalo:

$$x_i^* = \frac{LS_i + LI_i}{2}.$$

Frequência absoluta de um intervalo (f_i): número de observações pertencentes ao intervalo i.

Frequência relativa de um intervalo: $f_{ri} = f_i / n$.

Frequência acumulada absoluta de um intervalo:

$$F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j$$

Frequência acumulada relativa de um intervalo:

$$F_{r_i} = f_{r_1} + f_{r_2} + \cdots + f_{r_i} = \sum_{j=1}^{i} f_{r_j} \quad \text{ou} \quad F_{r_i} = \frac{F_i}{n}$$

Exemplo

Variável: viscosidade (em u.v.) de um líquido a uma certa temperatura.

```
13.9 14.9 15.9 15.8 14.8 15.1 15.8 15.0 15.1 14.6 14.7 16.6 13.6 15.9 13.1
15.2 14.7 16.0 15.6 17.4 15.3 14.2 15.9 15.1 15.9 16.1 16.2 13.8 14.6 16.0
15.8 15.5 16.5  17.1 15.3 15.5 17.8  15.4 15.4  14.6
```

Amostra ordenada:

```
13.1 13.6 13.8 13.9 14.2 14.6 14.6 14.6 14.7 14.7 14.8 14.9 15.0 15.1 15.1
15.1 15.2 15.3 15.3 15.4 15.4 15.5 15.5 15.6 15.8 15.8 15.8 15.9 15.9 15.9
15.9 16.0  16.0 16.1 16.2  16.5 16.6 17.1 17.4  17.8
```

```
n = 40
Mínimo  Mediana   Média  Máximo
 13.10    15.40   15.39   17.80
```

Procedimento:

Adotamos $k = 5$.

min = 13,10 e MAX = 17,80.

A = MAX – min = 17,8 – 13,10 = 4,7.

h = 4,7 / 5 = 0,94.

Adotamos $h = 1$ e $LI_1 = 13$.

Limites dos intervalos: $LI_1 = 13$, $LS_1 = LI_1 + h = 14$, $LI_2 = LS_1 = 14$, $LS_2 = LI_2 + h = 15$, …, $LI_5 = LS_4 = 17$ e $LS_5 = LI_5 + h = 18$.

Pontos médios: $x_1^* = \dfrac{13+14}{2} = 13{,}5; \quad x_2^* = \dfrac{14+15}{2} = 14{,}5; \quad \ldots; x_5^* = \dfrac{17+18}{2} = 17{,}5.$

Tabela. Distribuição de frequências da variável viscosidade.

Ordem	Classe	Ponto médio	Frequência	Frequência relativa	Frequência acumulada	Frequência relativa acumulada
1	13 |-- 14	13,5	4	0,1	4	0,1
2	14 |-- 15	14,5	8	0,2	12	0,3
3	15 |-- 16	15,5	19	0,475	31	0,775
4	16 |-- 17	16,5	6	0,15	37	0,925
5	17 |-- 18	17,5	3	0,075	40	1
		Total	40	1	-	-

Nesta organização de dados temos perda de informação.
Em um gráfico de pontos (lâmina 26) não há perda de informação, mas se n for “grande”, pode haver perda de clareza.

Densidade de frequência (ou densidade): $f_{d_i} = \dfrac{f_{r_i}}{h}.$

Representação gráfica:

Histograma

Gráfico de barras adjacentes com bases iguais às amplitudes dos intervalos e alturas iguais às densidades.

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

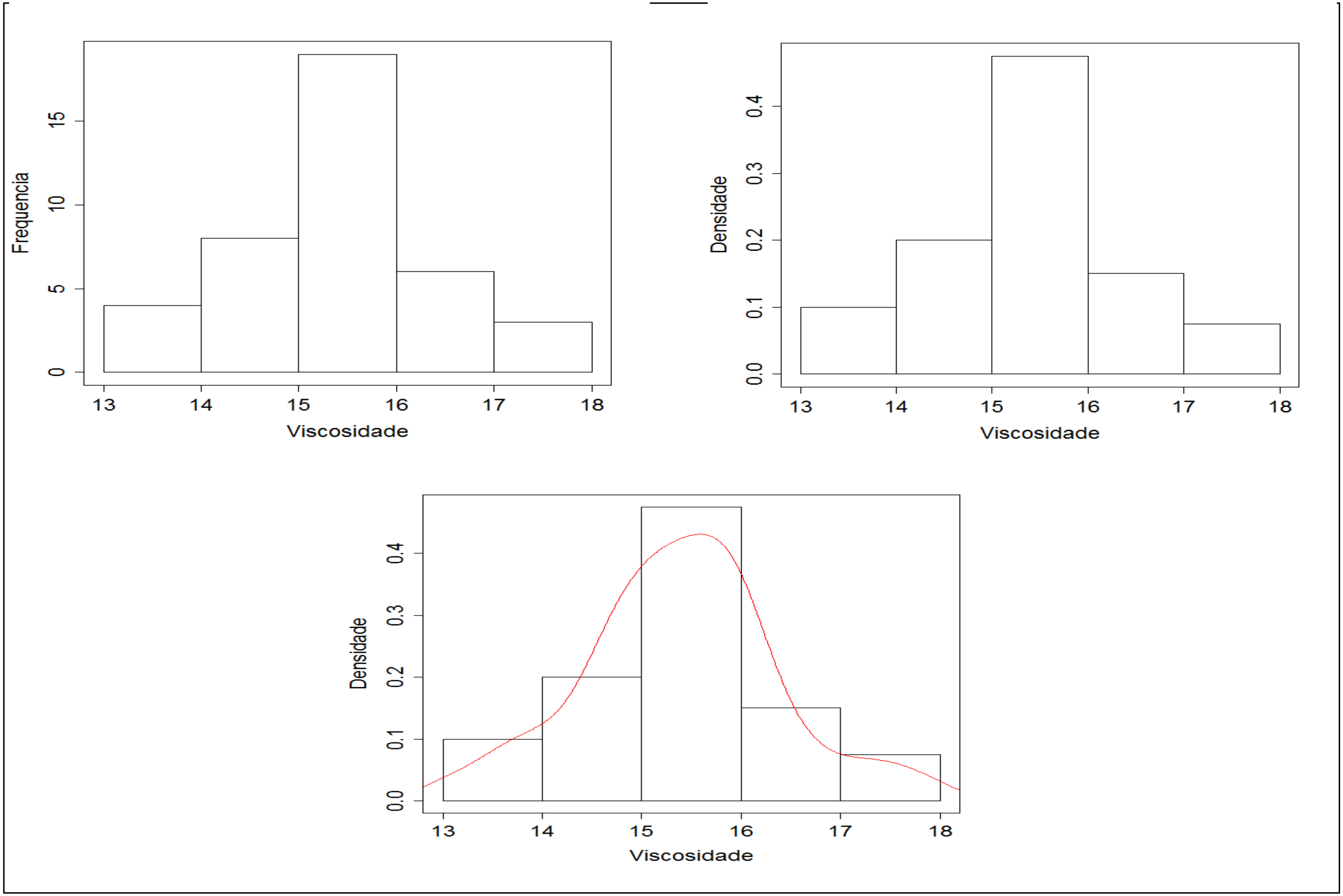
Propriedade. Se utilizarmos densidades, soma das áreas dos retângulos = 1, pois

$$\sum hf_{d_i} = \sum_{i=1}^{k} h \frac{f_{r_i}}{h} = \sum_{i=1}^{k} f_{r_i} = 1.$$

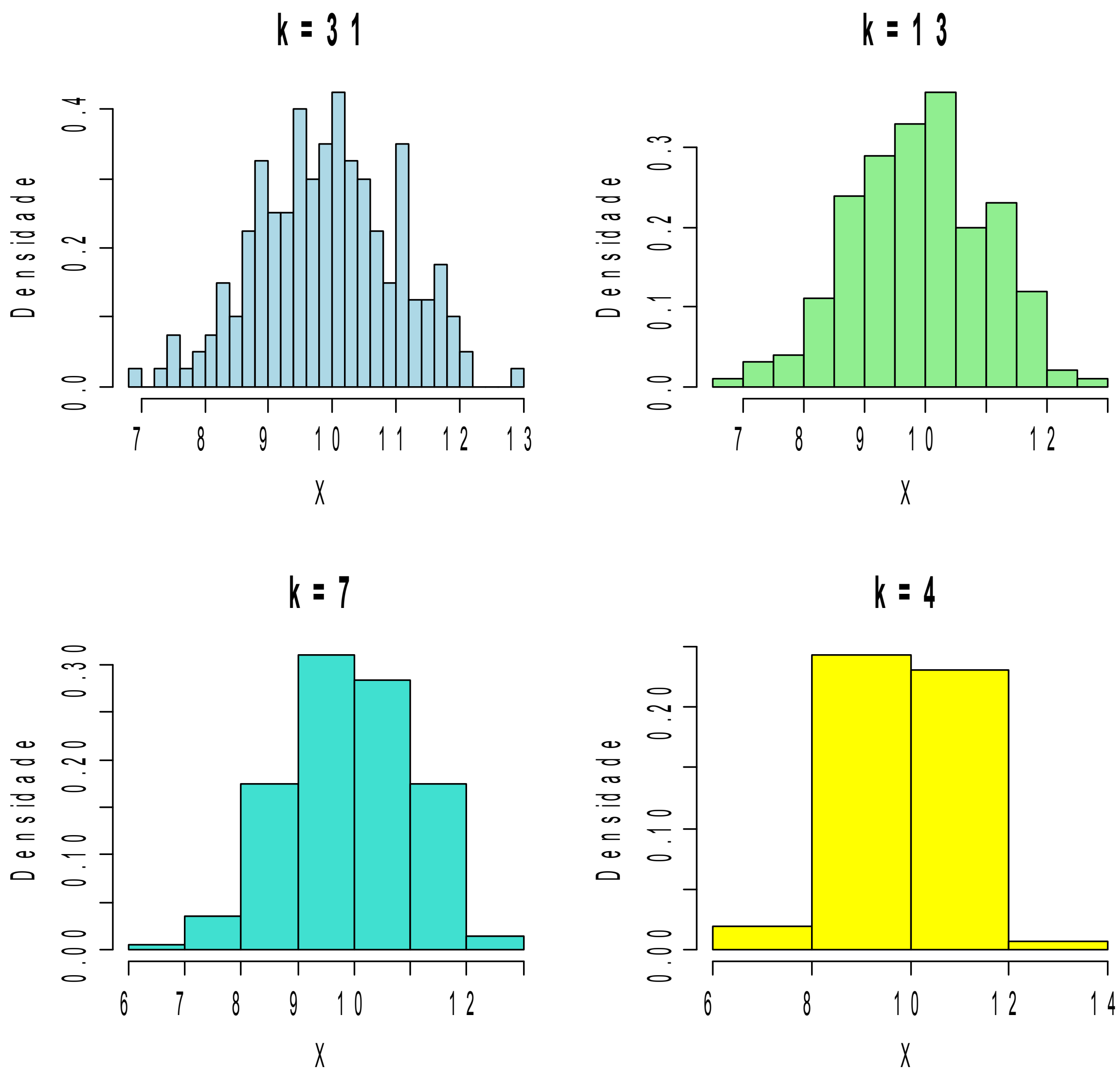
Obs. 1. A amplitude dos intervalos pode variar.

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

Exemplo. Variável viscosidade.



Escolha do número de classes (geralmente, $5 \leq k \leq 15$).



Média e variância para variáveis contínuas agrupadas em classes

Média:

$$\bar{x} \cong \frac{x_1^* f_1 + x_2^* f_2 + \cdots x_k^* f_k}{n} = \frac{\sum_{i=1}^{k} x_i^* f_i}{n}$$

Variância:

$$s^2 \cong \frac{\sum_{i=1}^{k} f_i \left(x_i^* - \bar{x}\right)^2}{n-1}$$

Exemplo. Variável viscosidade (lâmina 40)

$$\bar{x} \cong \frac{13{,}5 \times 4 + 14{,}5 \times 8 + 15{,}5 \times 19 + 16{,}5 \times 6 + 17{,}5 \times 3}{40}$$

$$= \frac{616}{40} = \quad 15{,}4.$$

$$s^2 \cong \frac{\sum_{i=1}^{5} f_i \left(x_i^* - \bar{x}\right)^2}{40-1} = \frac{41{,}6}{39} = 1{,}067.$$

$$\Rightarrow s = 1{,}033 \text{ (desvio padrão).}$$

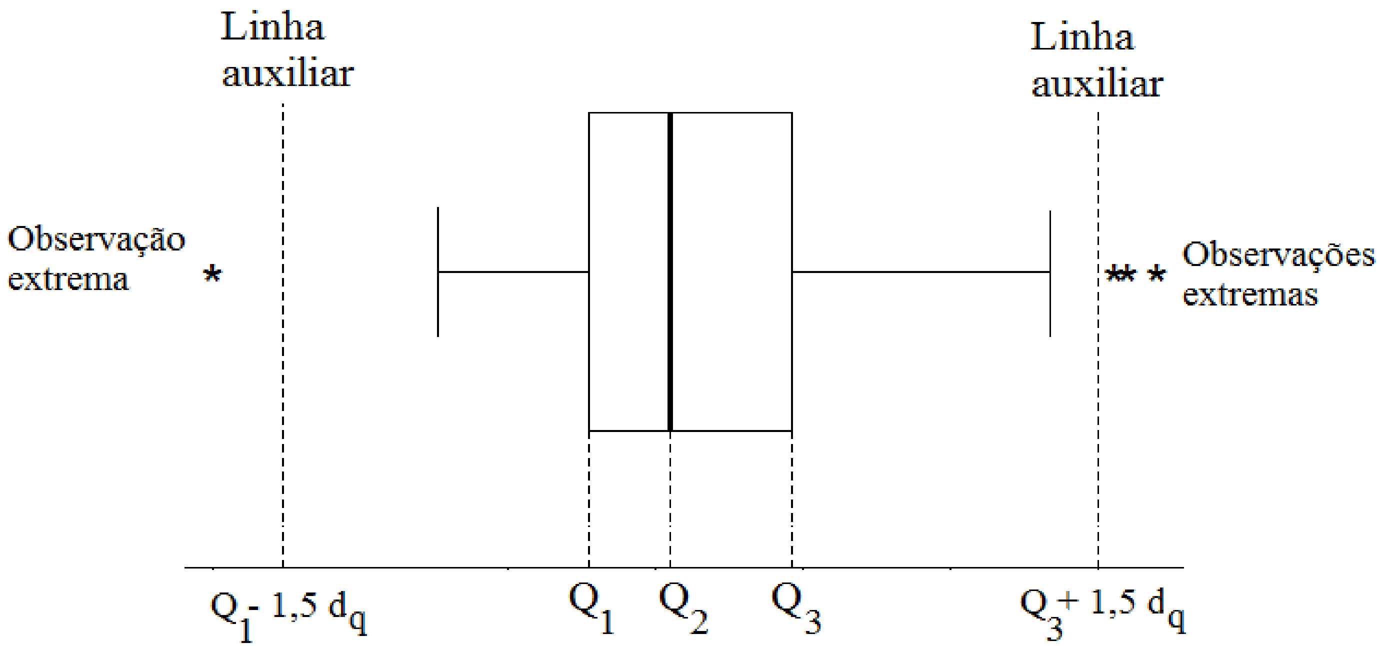
Média dos dados não agrupados (dados brutos) :

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{36}}{40} = \frac{13{,}9 + 14{,}9 + \cdots + 14{,}6}{40} = 15{,}39.$$

Este resultado difere do valor obtido anteriormente. Por quê?

Gráfico de caixas (*box plot* ou *boxplot*)

Representação dos dados por meio de um retângulo construído com os quartis. Fornece informação sobre a variabilidade ($d_q = Q_3 - Q_1$) e valores extremos.



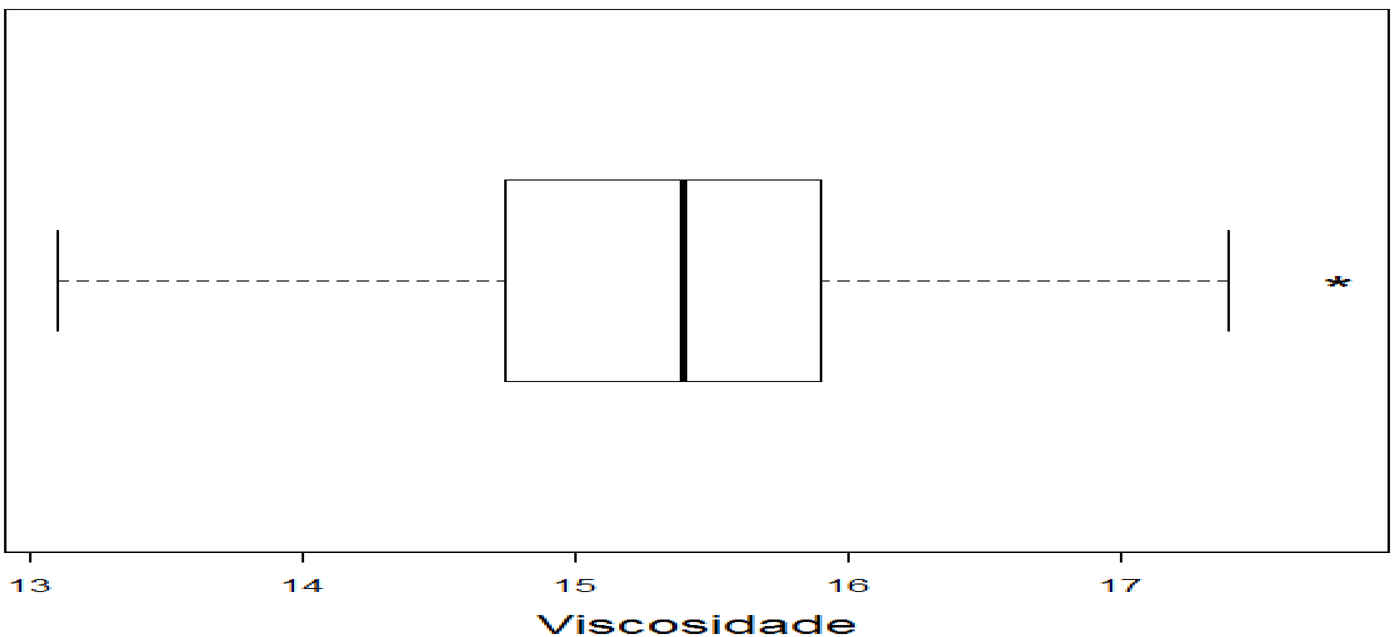
Vertical à esquerda ├ : menor valor na amostra que não é extremo.

Vertical à direita ┤ : maior valor na amostra que não é extremo.

Exemplo. Variável viscosidade.

1º quartil (Q1) = 14,775.
Mediana (Md ou Q2) = 15,4.
3º quartil (Q3) = 15,9.

d_q= intervalo interquartil = Q3 – Q1 = 1,125.
Lnhas auxiliares passam por Q1 – 1,5d_q = 13,1 e Q3 +1,5d_q = 17,6.



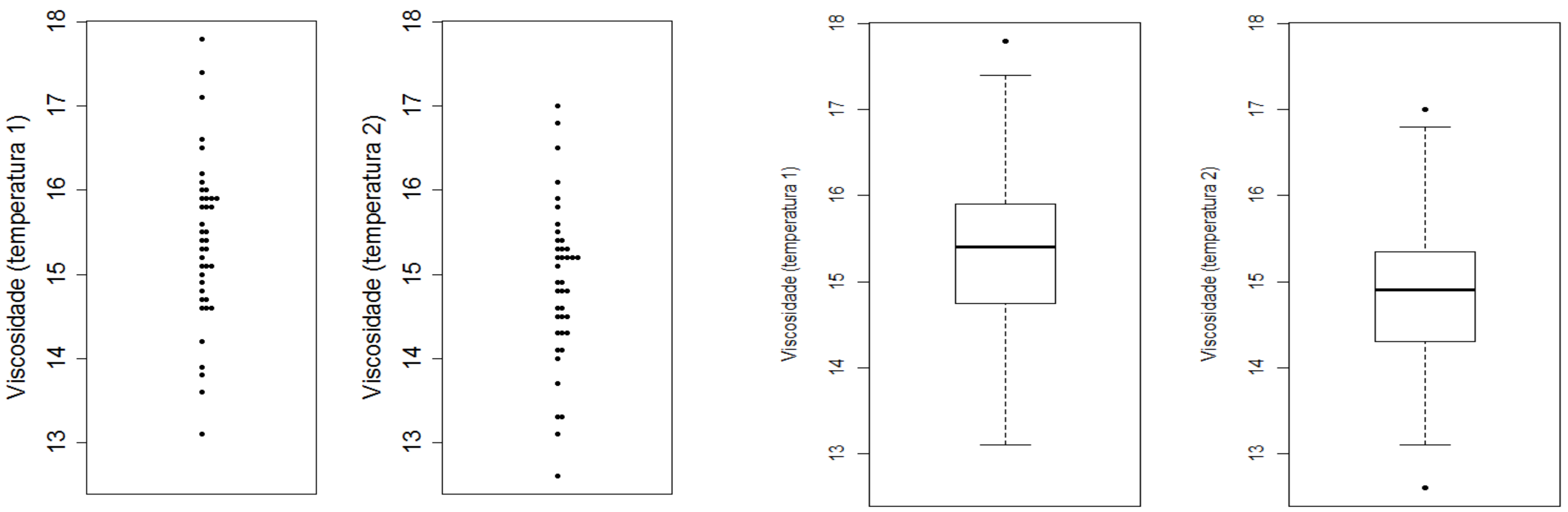
Exemplo. Variável viscosidade medida em duas temperaturas.

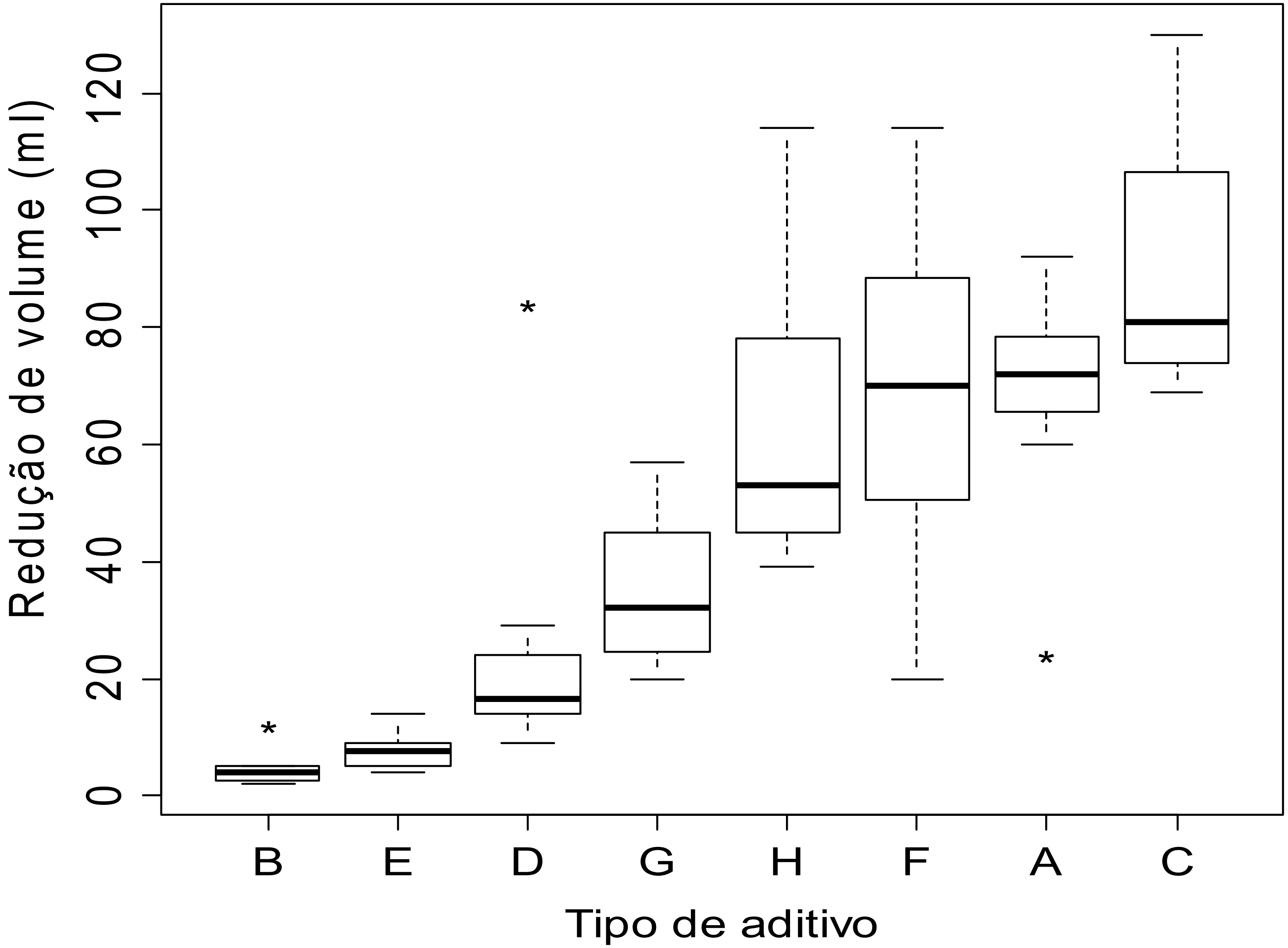
Temperatura 1 (lâmina 39).

13.9 14.9 15.9 15.8 14.8 15.1 15.8 15.0 15.1 14.6 14.7 16.6 13.6 15.9 13.1
15.2 14.7 16.0 15.6 17.4 15.3 14.2 15.9 15.1 15.9 16.1 16.2 13.8 14.6 16.0
15.8 15.5 16.5 17.1 15.3 15.5 17.8 15.4 15.4 14.6

Temperatura 2 (n = 40).

13.3 14.5 15.3 15.3 14.3 14.8 15.2 14.5 14.6 14.1 14.3 16.1 13.1 15.5 12.6
14.6 14.3 15.4 15.2 16.8 14.9 13.7 15.2 14.5 15.3 15.6 15.8 13.3 14.1 15.4
15.2 15.2 15.9 16.5 14.8 15.1 17.0 14.9 14.8 14.0

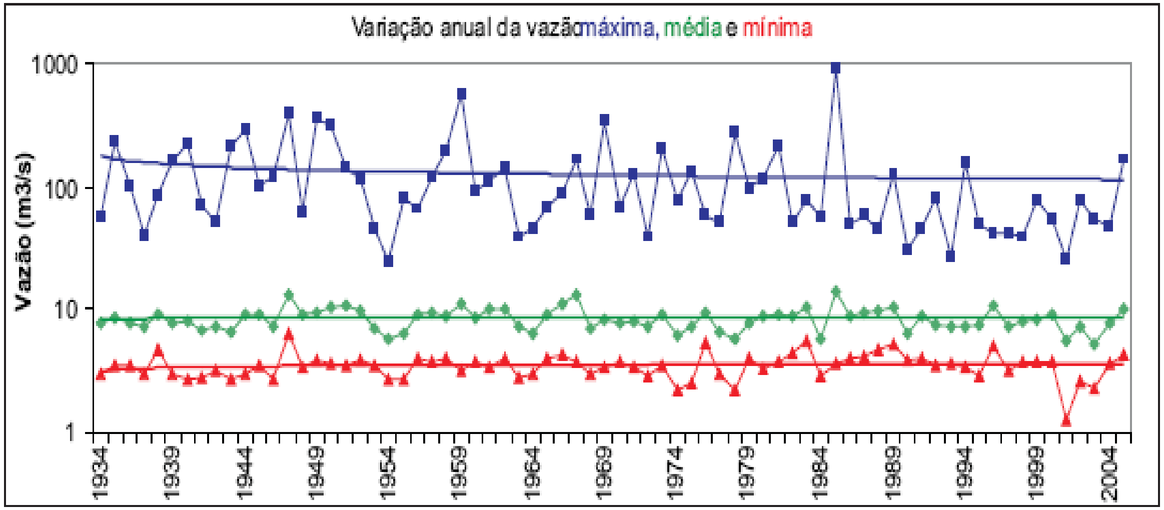


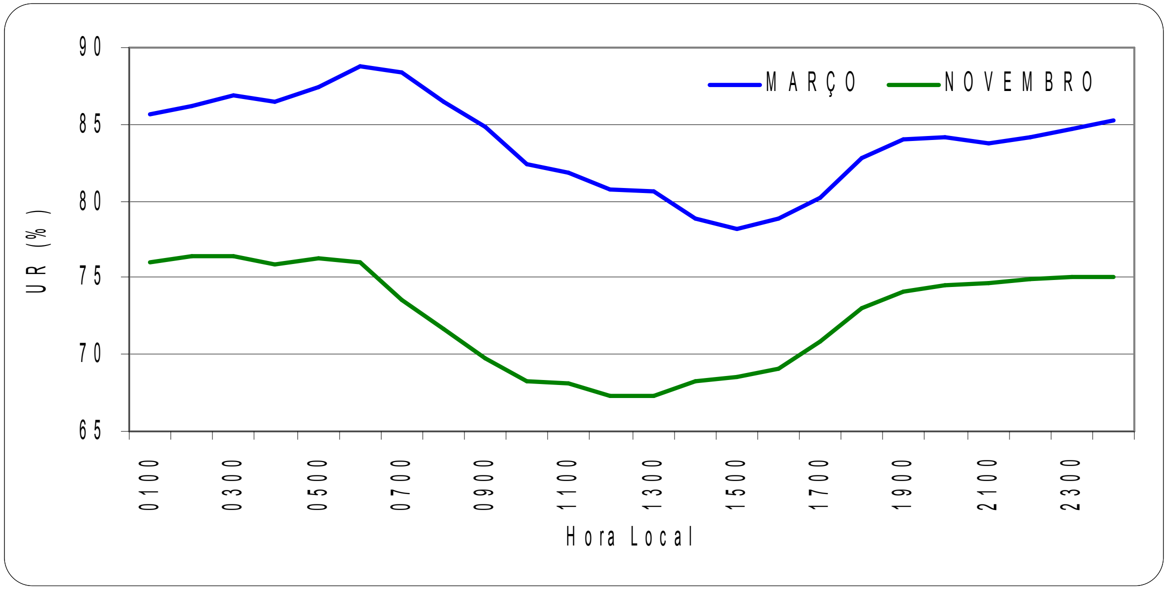


Análise exploratória. Redução de volume *versus* tipo de aditivo. Variabilidade. Simetria. Valores extremos.

Gráfico de linha

Representação de séries temporais (ou séries históricas).





Visualização dos componentes tendência e sazonalidade.

Cuidado com a escala do gráfico.

Associação entre variáveis quantitativas

$(x_1,y_1), \ldots, (x_n,y_n)$: amostra bivariada.

Representação gráfica: gráfico de dispersão (*scatter plot*)

Medida de associação: coeficiente de correlação linear de Pearson.

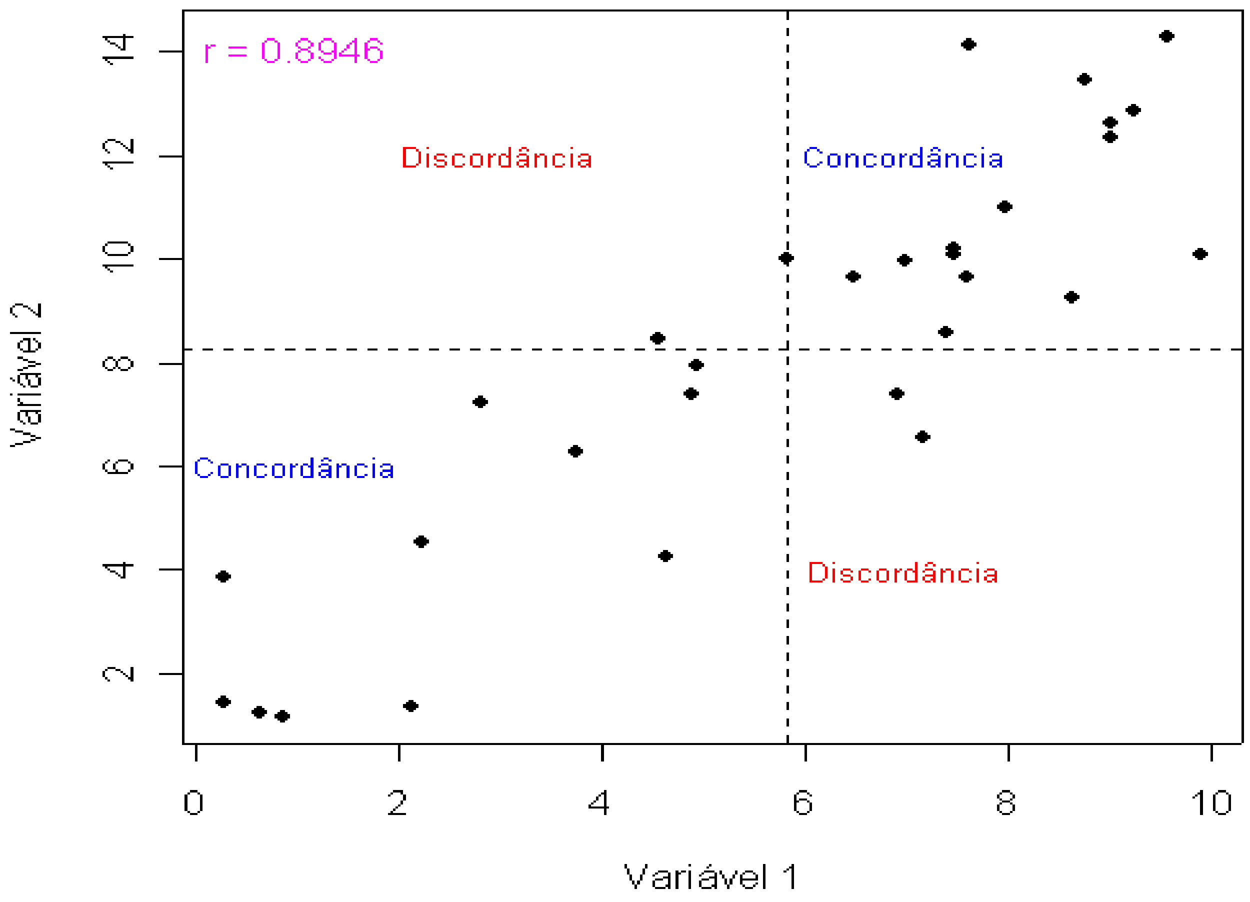
$$r = \frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

Numerador: covariância entre x e y.

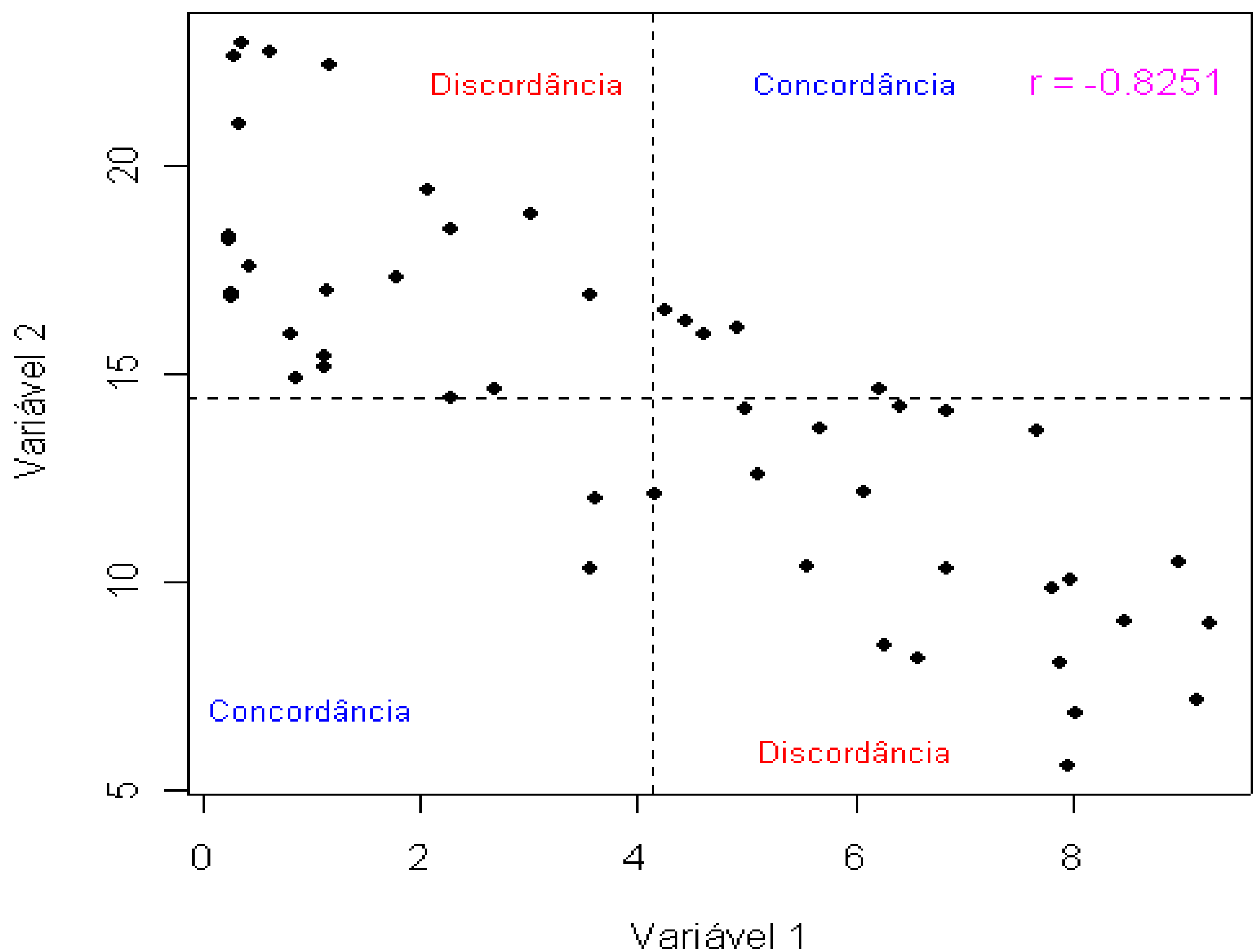
Propriedades: (1) $-1 \leq r \leq 1$ e

(2) $|r| = 1$ se, e somente se, a relação entre x e y for linear ($y = a + bx$, $b \neq 0$ e o sinal de r é o sinal de b).

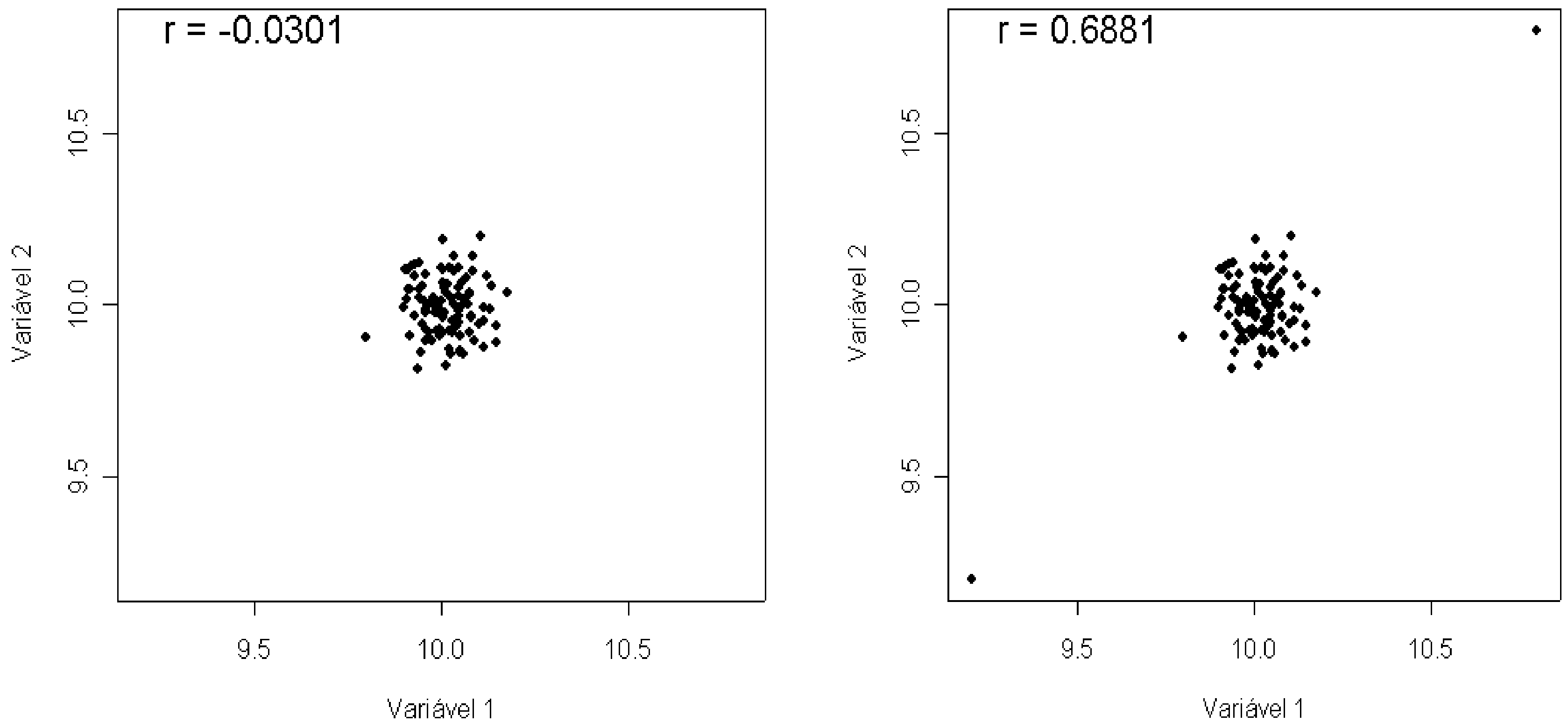
Associação entre variáveis quantitativas



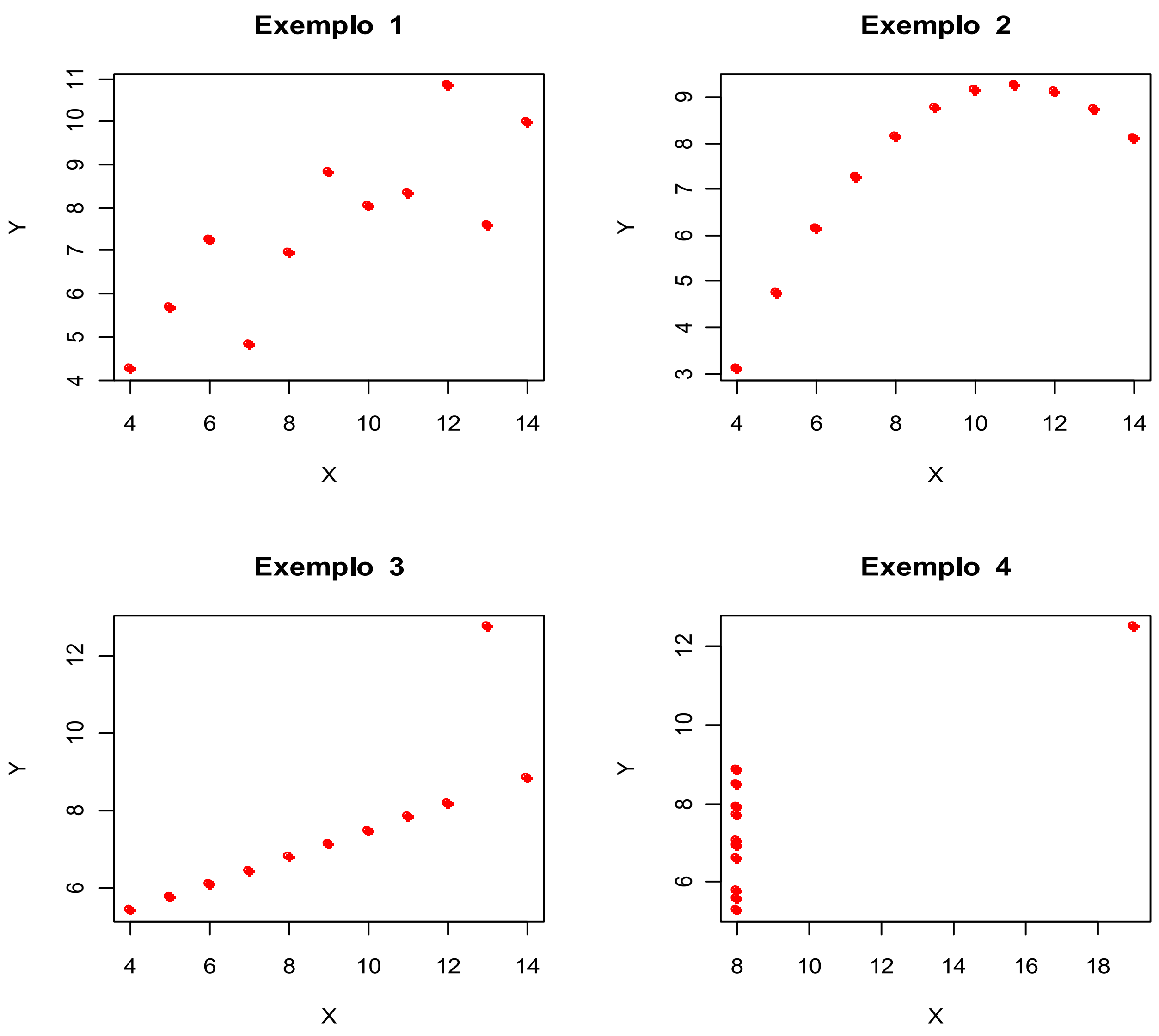
Associação entre variáveis quantitativas



Associação entre variáveis quantitativas



Associação entre variáveis quantitativas



Correlações:

Exemplo 1: 0,8164

Exemplo 2: 0,8162

Exemplo 3: 0,8163

Exemplo 4: 0,8165