Conceitos de estatística descritiva

A estatística descritiva pode ser entendida como uma ferramenta capaz de descrever ou resumir dados, mostrando aspectos importantes do conjunto de dados, como o tipo de distribuição associada e os valores mais representativos do conjunto, e permitindo criar visualizações referentes a tais aspectos.

População, amostra e variáveis

Os conceitos de população , amostra e variáveis são brevemente comentados nesta seção.

População

População, ou universo, é o nome que se dá a um conjunto de unidades (elementos ou exemplares/dados se aproximarmos a definição ao contexto de mineração de dados) que compartilham características comuns e sobre o qual é pretendido o desenvolvimento de um estudo.

O conjunto de vendas realizadas durante o tempo de funcionamento de um restaurante é um exemplo de população (de vendas).

A população pode ser finita ou infinita.

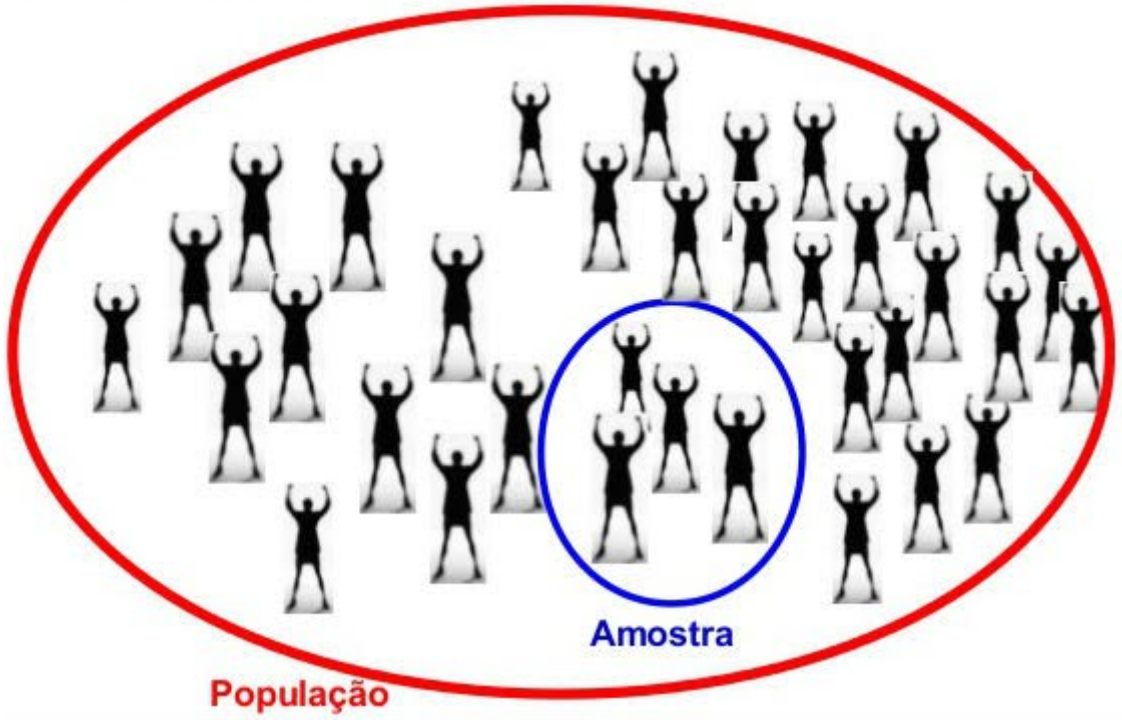
Sendo finita, se pequena, um estudo pode envolver sua totalidade. Caso seja muito grande ou infinita, faz-se necessário estabelecer uma amostra, para que um estudo sobre ela se viabilize.

Amostra

Uma amostra, então, é um subconjunto da população (também chamado de fração ou parte) que deve ser estabelecido com cuidado, seguindo técnicas apropriadas, pois o objetivo de obter uma amostra é reduzir o tamanho da população sem que características essenciais associadas a ela sejam perdidas.

Boas técnicas de amostragem geram amostras representativas e imparciais da população, ou seja, mantêm a proporcionalidade dos fenômenos que ocorrem na população e conferem chances iguais aos elementos da população de fazerem parte da amostra.

Há diferentes formas de estabelecer uma amostra, que resultam em diferentes tipos de amostragem: aleatória, estratificada, por conglomerados, acidental, intencional, por quotas etc...


Amostra
População

Variável

Na estatística descritiva, uma variável diz respeito a uma característica associada a elementos de uma população.

Por exemplo, num restaurante, uma variável pode ser a temperatura em que um vinho deve ser servido (com valores como 5°C ou 12°C) ou o tipo de um prato presente no cardápio (com os seguintes valores, por exemplo: aperitivo, pasta, carne branca ou sobremesa quente).

Fazendo uma analogia à representação de um dado usada neste livro, uma variável é equivalente a um atributo descritivo. As variáveis, no contexto da estatística descritiva, podem ser **quantitativas** ou **qualitativas**.


TIPOS DE VARIÁVEIS
QUANTITATIVAS
DISCRETAS
0 1 2 3
CONTÍNUAS
0,0 1,0 2,0 3,0
QUALITATIVAS
CATEGÓRICAS
ORDINAL
NOMINAL
DICOTÔMICAS
SIM
NÃO
YOUTUBE.COM/C/ROGERIOANTON

Variável Quantitativa

As **variáveis quantitativas** (também nomeadas no contexto de mineração de dados como variáveis numéricas) são aquelas que podem ser medidas em uma escala de valores numéricos, e se dividem em **discretas** (assumem valores dentro de um conjunto finito ou infinito contável) ou **contínuas** (assumem valores em uma escala contínua).

Variável Qualitativa

As **variáveis qualitativas** (também nomeadas no contexto de mineração de dados como variáveis categóricas) são aquelas que representam uma classificação do elemento ao qual o valor da variável está associado, e se dividem em **nominais** (quando não há ordenação entre valores) ou **ordinais** (quando há uma ordenação entre os valores). Alguns exemplos de variáveis, considerando o contexto de um restaurante, e alguns valores que elas podem assumir estão organizados a seguir

Variáveis quantitativas

Discretas

Quantidade de bebidas vendidas: 15, 6, 3, 0
Idade do cliente: 5, 10, 25, 32, 57

Contínuas

Preço de um item do cardápio: 5,50, 35,00, 99,50
Peso de um corte de carne: 100g, 226g, 500g

Variáveis qualitativas

Nominais

Código da forma de pagamento: 11, 21, 31
Forma de pagamento: dinheiro, débito, crédito

Ordinais

Faixa etária dos clientes: 0 a 10 anos, 11 a 17 anos, 18 a 30 anos, 31 a 50 anos, mais de 50 anos
Mês de observação: janeiro, abril, novembro

Medidas de posição e separatrizes

Medidas de posição , ou medidas de tendência central , permitem encontrar os valores que orientam a análise dos dados no que diz respeito à sua localização, ou como a distribuição associada aos valores se comporta no universo amostrado.

As medidas de posição mais comuns no contexto de análise exploratória de dados são média aritmética, mediana e moda. Relacionadas com a mediana estão as medidas separatrizes percentis e quartis.

Para um conjunto de valores $v = v_1, v_2, \ldots, v_n$, em que n é a quantidade de valores do conjunto, a média aritmética simples (média(v) ou) é a soma dos valores do conjunto v, dividida pela quantidade de valores presentes nesse conjunto.

Formalmente, $m\acute{e}dia(v)$ é definida como:

$$media(v) = \frac{1}{n} \sum_{i=1}^{n} v_i$$

A **mediana** ($mediana(v)$) é o valor que divide a distribuição dos valores exatamente ao meio (podendo ser vista também como uma medida separatriz).

Importante lembrar que tal valor não precisa estar presente no conjunto v.

Para o cálculo da mediana, todos os valores presentes no conjunto devem ser ordenados de forma crescente (formando o conjunto v'), e a mediana será, então, dada por:

$$mediana\,(\mathbf{v}) = \begin{cases} v'_i & \text{se } n \text{ é impar, } i = \dfrac{n+1}{2} \\ \dfrac{1}{2}(v'_i + v'_{i+1}) & \text{se } n \text{ é par, } \quad i = \dfrac{n}{2} \end{cases}$$

em que i é uma posição no conjunto de valores v'.

A **moda** , por sua vez, é o valor mais frequente em um conjunto de valores.

Ela é a única medida de posição que pode assumir mais de um valor.

Essa situação ocorre quando dois ou mais valores aparecem no conjunto de valores v com a mesma frequência, a máxima no conjunto. Assim, um conjunto de valores pode ser:

- **amodal** (não possui moda),
- **unimodal** (possui uma moda),
- **bimodal** (possui duas modas) ou
- **multimodal** (possui diversas modas).

Exemplos

Para exemplificar o uso dessas medidas e como a comparação entre elas pode ser útil, considere os dados da Tabela 1 a seguir, que representa a quantidade de vendas, durante um mês, dos nove itens presentes no cardápio promocional de um restaurante.

Quantidade de vendas de nove itens do restaurante durante um mês

Item	712	68	2	65	103	809	111	601	44
Quantidade	29	30	32	65	65	65	25	25	90

Presumindo que se queira calcular as medidas de posição em relação à variável **Quantidade**, por causa da **mediana**, e apenas por causa dela, é preciso ordenar os dados da Tabela 2 considerando os valores dessa variável (Tabela 2).

Quantidade de vendas ordenada

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Item	601	111	712	68	2	65	103	809	44
Quantidade	25	25	29	30	32	65	65	65	90

Item	712	68	2	65	103	809	111	601	44
Quantidade	29	30	32	65	65	65	25	25	90

Iniciando a análise exploratória desses dados pelo cálculo da média, com o uso da Fórmula da Média, tem-se:

$$m\acute{e}dia(Quantidade) = (25 + 25 + 29 + 30 + 32 + 65 + 65 + 65 + 90)/9 = 47,3.$$

Para o cálculo da mediana, desde que haja uma quantidade ímpar de valores, tem-se:

$$i = (n + 1)/2 = (9 + 1)/2 = 5$$

e, portanto, $mediana(Quantidade) = 32,$

ou seja, a mediana é o valor da variável Quantidade na quinta posição da ordenação.

Enfim, a moda é o valor com maior quantidade de repetições.

No caso da variável Quantidade, o valor 65 aparece três vezes, enquanto todos os demais valores aparecem uma ou duas vezes; portanto, 65 é a moda do conjunto.

As medidas de posição calculadas para a variável Quantidade estão resumidas na Tabela 3

Medidas de posição para a variável Quantidade

Média	Mediana	Moda
47,33	32	65

Em termos práticos, essas medidas permitem algumas intepretações sobre o cardápio:

- durante o mês de observação, a quantidade média de vendas de um item foi de 47,33 unidades;
- a mediana revela que, embora as vendas para o item 002 sejam em quantidade maior que as vendas para metade dos itens observados (é o quinto mais vendido dentro de nove itens), a venda desse item está abaixo da média de vendas do conjunto.

Entendendo um pouco masi sobre os dados de vendas

A comparação da média, mediana e moda pode revelar informações interessantes sobre a distribuição dos valores do conjunto em questão.

Quando essas medidas assumem o mesmo valor (ou valores com variações muito pequenas entre si) significa que o conjunto de valores de uma variável tem **simetria**, ou seja, é possível usar uma dessas medidas de posição para separar o conjunto de valores em duas partes com a mesma distribuição de frequência.

Por outro lado, se os valores são diferentes, diz-se que a distribuição dos valores da variável é **assimétrica** .

No exemplo da variável Quantidade, a média, a mediana e a moda assumem valores diferentes, o que indica que a distribuição da quantidade de vendas é assimétrica.

Nesse caso, menos da metade dos itens vende mais que a média (quatro itens vendem mais e cinco itens vendem menos).

Percentis e Quartis

Os **percentis** e **quartis** são medidas separatrizes que dividem o conjunto de valores, ordenado de forma crescente, em partes tão iguais quanto possível.

O **percentil de ordem p** determina os **p% menores valores contidos em v'**, e a posição i, que delimita o percentil de ordem p em v', é dada por:

$$i = \frac{p(n+1)}{100}$$

Isso significa que os p% menores valores em v' estão abaixo do valor da posição i.

Então, considerando a variável Quantidade (Tabela 2), o percentil de ordem p = 50 indica os 50% menores valores no conjunto ordenado e estão abaixo do valor localizado na posição dadda por:

$$i = \frac{50(9+1)}{100} = 5$$

Assim, o valor 32 é o percentil de ordem 50 em v'.

	1	2	3	4	5	6	7	8	9
Quantidade	25	25	29	30	32	65	65	65	90

Casos particulares de percentis definem os quartis

O primeiro quartil , ou Q1, é o percentil de ordem p = 25, ou seja, o valor de v que separa os 25% valores menores que vi dos 75% valores maiores que vi, sendo vi o valor que ocupa a $i = \frac{25(n+1)}{100}$

O segundo quartil, ou Q2, é o percentil de ordem p = 50 e é equivalente à mediana do conjunto de valores.

O terceiro quartil, ou Q3, é equivalente ao percentil de ordem p = 75 e é o valor que separa os 75% valores menores que vi dos 25% valores maiores que v_i, sendo v_i o valor que ocupa a $i = \frac{75(n+1)}{100}$ posição em v'.

Para o caso da variável Quantidade (Tabela 2), Q1, Q2 e Q3 são, respectivamente,

- $v_3 \quad , \\ = 29$
- $v_5 \quad , \\ = 32$
- $v_8 \quad . \\ = 65$

Em termos práticos, poderíamos dizer que o item 044 está acima do percentil de ordem 75 ou acima de Q3, indicando que pelo menos 75% dos itens observados vendem menos que ele, evidenciando sua boa aceitação por parte dos clientes.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Item	601	111	712	68	2	65	103	809	44
Quantidade	25	25	29	30	32	65	65	65	90

Medidas de dispersão

As medidas de posição são úteis para apresentar uma sumarização dos dados. No entanto, não são capazes de descrever a variação ou dispersão do conjunto de valores.

Para esse fim, aplicam-se as medidas de dispersão , capazes de descrever o quanto os valores de um conjunto estão próximos ou distantes de uma medida central, como a média.

As medidas mais comuns de dispersão ou variância dos dados são a **amplitude**, **variância**, **desvio-padrão**.

Amplitude

Para um conjunto de valores v, a medida da amplitude é a diferença entre o maior e o menor valor do conjunto, portanto:

$$amplitude(v) = max(v) - min(v).$$

Embora a medida da amplitude seja simples, sua interpretação precisa ser feita com cuidado, pois ela pode ser influenciada por valores extremos, conhecidos como outliers que veremos mais a frente.

Variância

A **variância** $\sigma^2(v)$ é uma medida de dispersão definida como a média dos quadrados das diferenças entre cada valor do conjunto v e a média desse conjunto. Formalmente, tem-se:

$$\sigma^2(v) = \frac{1}{n}\sum_{i=1}^{n}(v_i - media(v))^2$$

Desvio-padrão

O **desvio-padrão** $\sigma(v)$ é a raiz quadrada da variância.

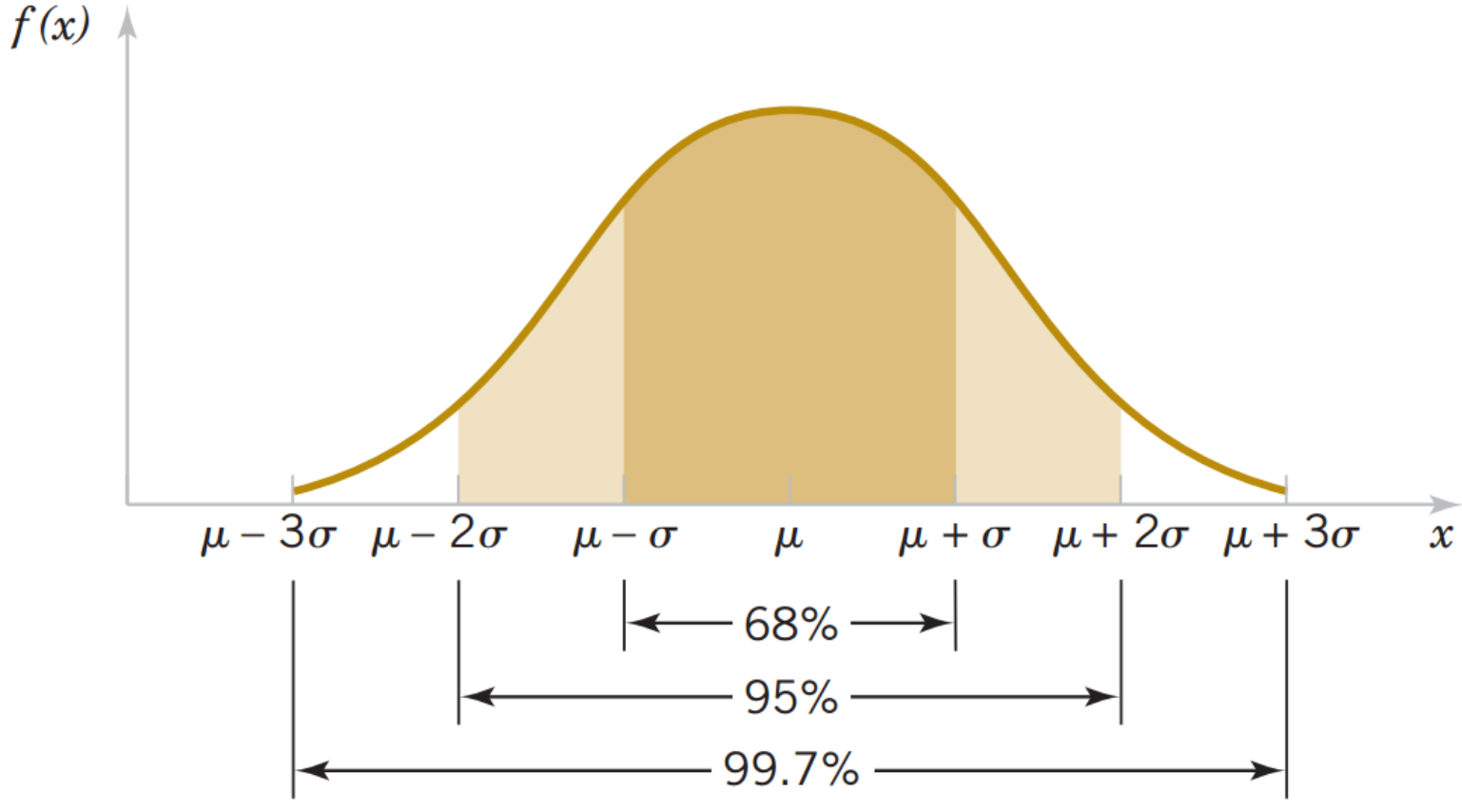
O uso do desviopadrão só faz sentido quando a média for usada como medida de posição (ou de centro).

Ele é semelhante à medida de amplitude, com a diferença de que o cálculo do desvio-padrão usa todos os valores de um conjunto.

O resultado dessa medida geralmente é usado para verificar a consistência de um fenômeno (um fenômeno é consistente quando o cálculo do desvio-padrão resulta em valores baixos).

Outra utilidade para o desvio-padrão é conhecida como regra empírica , ou regra 68-95-99 ou ainda regra dos 3-sigmas .

- Essa regra se aplica a conjunto de valores com distribuição normal e afirma que aproximadamente 68% dos valores desse conjunto estão a menos de um desvio-padrão da sua média;
- 95% dos valores desse conjunto estão a menos de dois desvios-padrão da sua média;
- 99,7% dos valores desse conjunto estão a menos de três desvios-padrão da sua média.


f(x)
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ
x
68%
95%
99.7%

Exemplo Utilizando o R

In [105]:
```
Quantidade <- c(29,30,32,65,65,65,25,25,90)
```

In [106]:
```
Quantidade
```

29 · 30 · 32 · 65 · 65 · 65 · 25 · 25 · 90

In [107]:
```
mean(Quantidade)
```

47.3333333333333

In [108]:
```
median(Quantidade)
```

32

In [109]:
```
Mode <- function(x) {
     ux <- unique(x)
     ux[which.max(tabulate(match(x, ux)))]
}
Mode(Quantidade)
```

65

In [110]:
```
sd(Quantidade)
```

24.0468293128221

In [111]:
```
var(Quantidade)
```

578.25

In [112]:
```
range(Quantidade)
```

25 · 90

In [113]:
```
max(Quantidade)
```

90

In [114]:
```
min(Quantidade)
```

25

In [115]:
```
amplitude = max(Quantidade)-min(Quantidade)
amplitude
```

65

In [116]:
```
Q1 <- quantile(Quantidade, probs = 0.25)
Q2 <- quantile(Quantidade, probs = 0.50)
Q3 <- quantile(Quantidade, probs = 0.75)
Q4 <- quantile(Quantidade, probs = 1.00)

quartis <- c(Q1,Q2,Q3,Q4)

quartis
```

25%: 29 **50%:** 32 **75%:** 65 **100%:** 90

In [117]:
```
Q1
```

25%: 29

Exemplo Prático de Nálise do dataset Iris

Como não poderia deixar de ser, a primeira parte de um projeto de data science segue a mesma lógica de um projeto de análise estatística de dados. Precisamos fazer a análise exploratória a fim de entender o conjunto de dados com o qual estamos trabalhando.

O conjunto de dados que vou baixar se chama Iris Flower Dataset. Por ser multivariado e com um número razoável de observações, este conjunto é bastante famoso na literatura estatística. Até mesmo Fisher já trabalhou conceitos de análise multivariada com estes dados.

Seu nome é Iris porque foram coletadas observações de 150 sujeitos de 3 espécies de flores do gênero Iris: Iris setosa, Iris versicolor e Iris virginica. Mais informações sobre estas flores podem ser encontradas na Wikipedia (https://en.wikipedia.org/wiki/Iris_flower_data_set).

iris setosa

iris versicolor

iris virginica

In [118]:

```r
# definir o endereco do conjunto de dados e baixa-lo
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
iris <- read.csv(url, header = FALSE)
names(iris) <- c("c_sepala", "l_sepala", "c_petala", "l_petala",
    "especie")
# remover a string 'Iris-' do inicio de cada tipo de especie
iris$especie <- as.factor(gsub("Iris-", "", iris$especie))
```

Com isso, percebemos que temos um conjunto de dados formado por cinco variáveis. Quatro destas variáveis são quantitativas e uma é categórica. As variáveis são

- c_sepala: comprimento da sépala das flores
- l_sepala: largura da sépala das flores
- c_petala: comprimento da pétala das flores
- l_petala: largura da pétala das flores
- especie: espécie da flor

In [119]:
```
head(iris) # exibe as 6 primeiras linhas do conjunto de dados para me dar uma ideia do q
ue estou enfrentando
```

A data.frame: 6 × 5

	c_sepala	l_sepala	c_petala	l_petala	especie
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

In [120]:
```
names(iris) # Exibe as colunas do dataframe iris
```

'c_sepala' · 'l_sepala' · 'c_petala' · 'l_petala' · 'especie'

In [121]:
```
# O resumo summary apresenta alguns detalhes muito importantes, incluindo percepções est
atísticas
summary(iris)
```

```
   c_sepala        l_sepala        c_petala        l_petala
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
      especie
setosa    :50
versicolor:50
virginica :50
```

In [122]:
```
sd(iris$c_sepala)
```

0.828066127977863

In [123]:
```
var(iris$c_sepala)
```

0.685693512304251

In [124]:
```
range(iris[,1:4])
```

0.1 · 7.9

In [125]:
```
head(iris[,1:4])
```

A data.frame: 6 × 4

	c_sepala	l_sepala	c_petala	l_petala
	<dbl>	<dbl>	<dbl>	<dbl>
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4

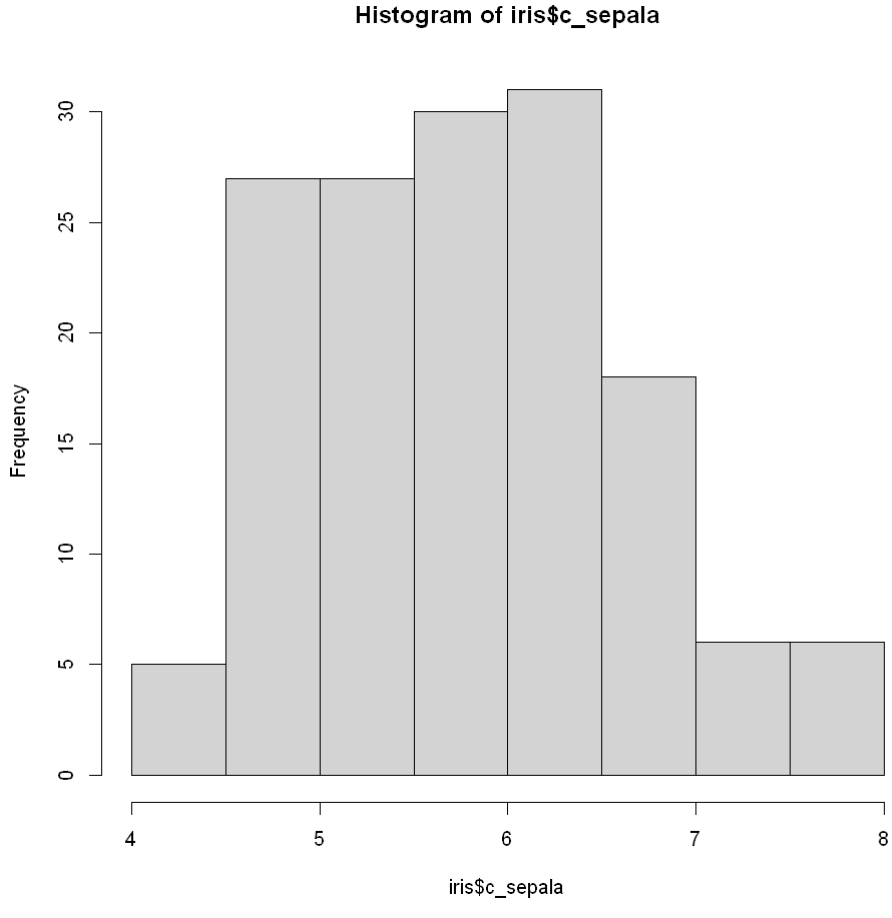
ESTATÍSTICAS DESCRITIVAS

Qual é a distribuição quantitativa das variáveis?

Esta pergunta é respondida usando um gráfico do histograma e boxplot das variáveis.

In [126]:

```
library(repr)
options(repr.plot.width=8, repr.plot.height=8)
hist(iris$c_sepala)
```

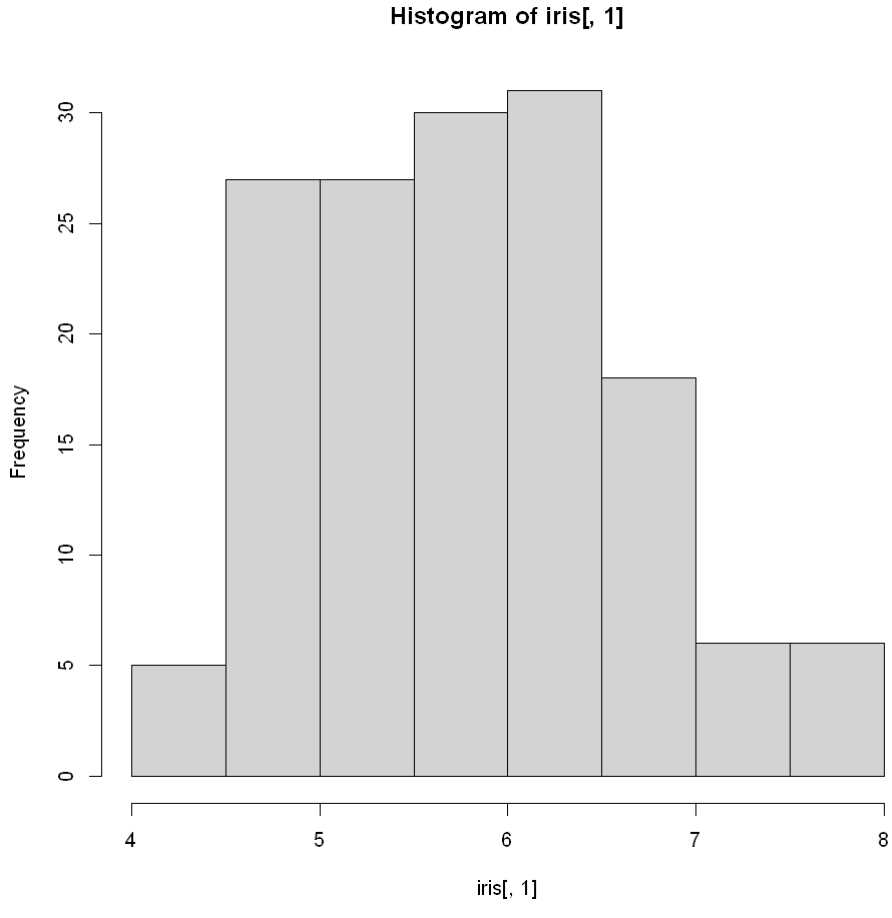


Alguns insights do histograma:

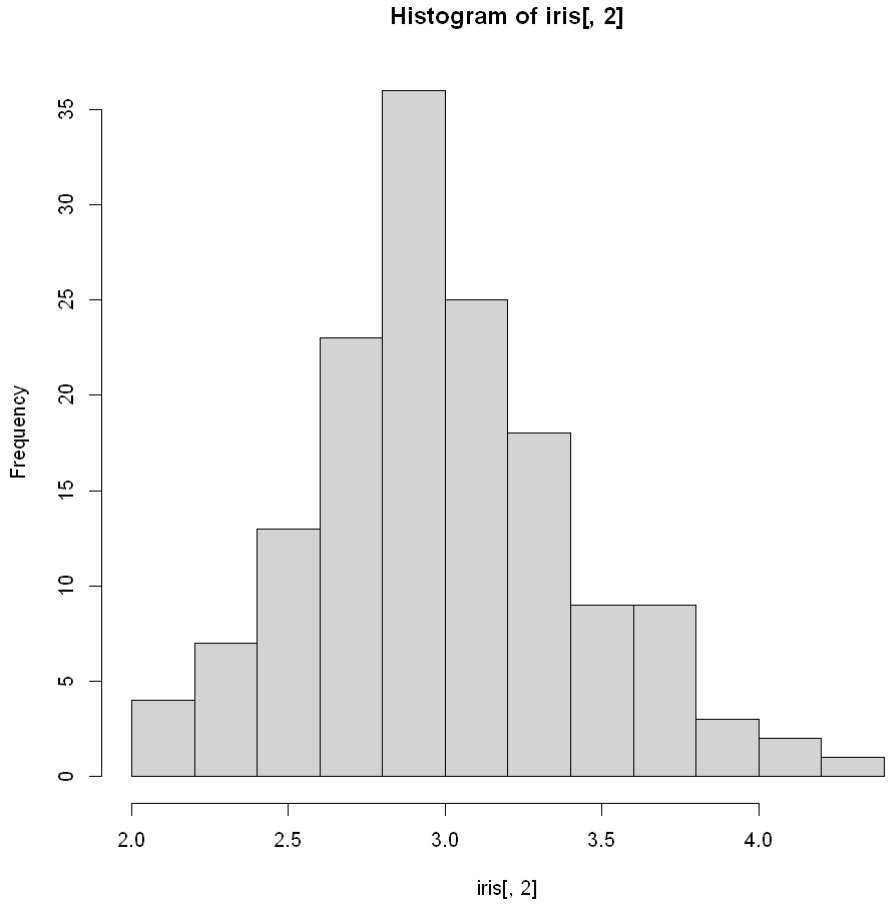
1. A partir do histograma, vemos que o comprimento da sépala varia de 4 a 8
2. enquanto a frequência varia de 5 a 30.
3. Também pode ser visto que a frequência mais alta tem comprimento de sépala que varia de 6 a 6,5 enquanto a segunda frequência mais alta tem uma faixa de comprimento sépala entre 5,5 e aproximadamente 6
4. O comprimento da sépala com a menor frequência varia de 4 a 4,5

A questão de saber se os dados são normalmente distribuídos ou não é abordada muito mais tarde, mas esses dados parece que está inclinado para a esquerda: isso será explorado mais a fundo ...

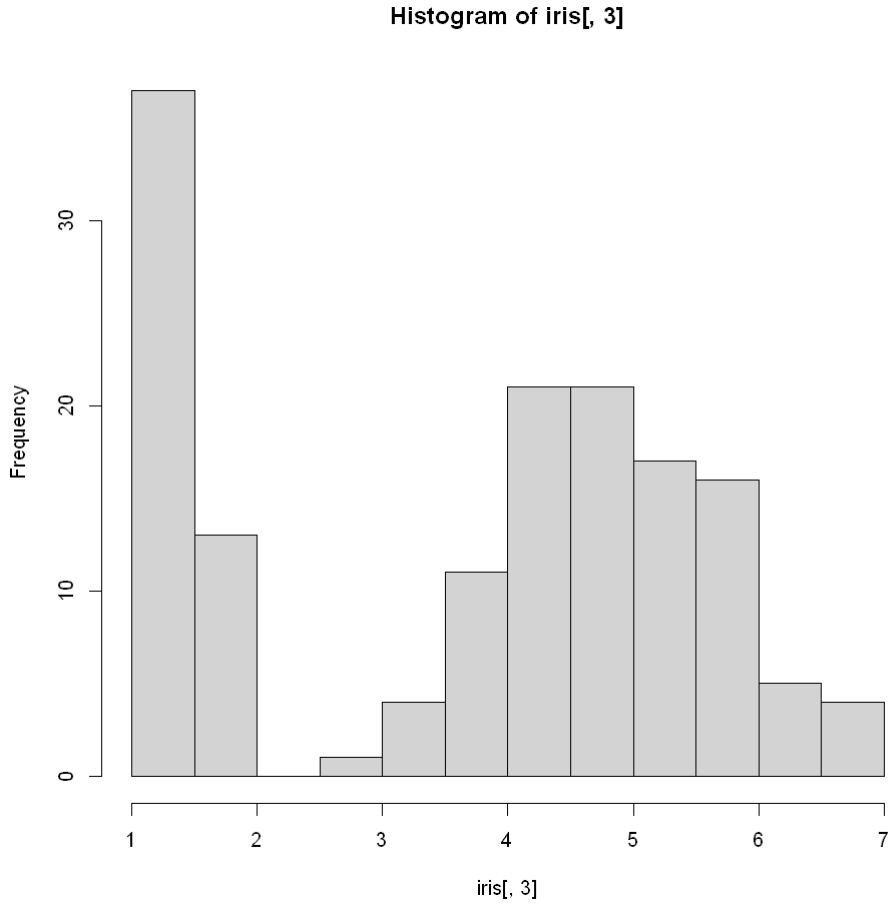
In [127]:
```
hist(iris[,1])
```



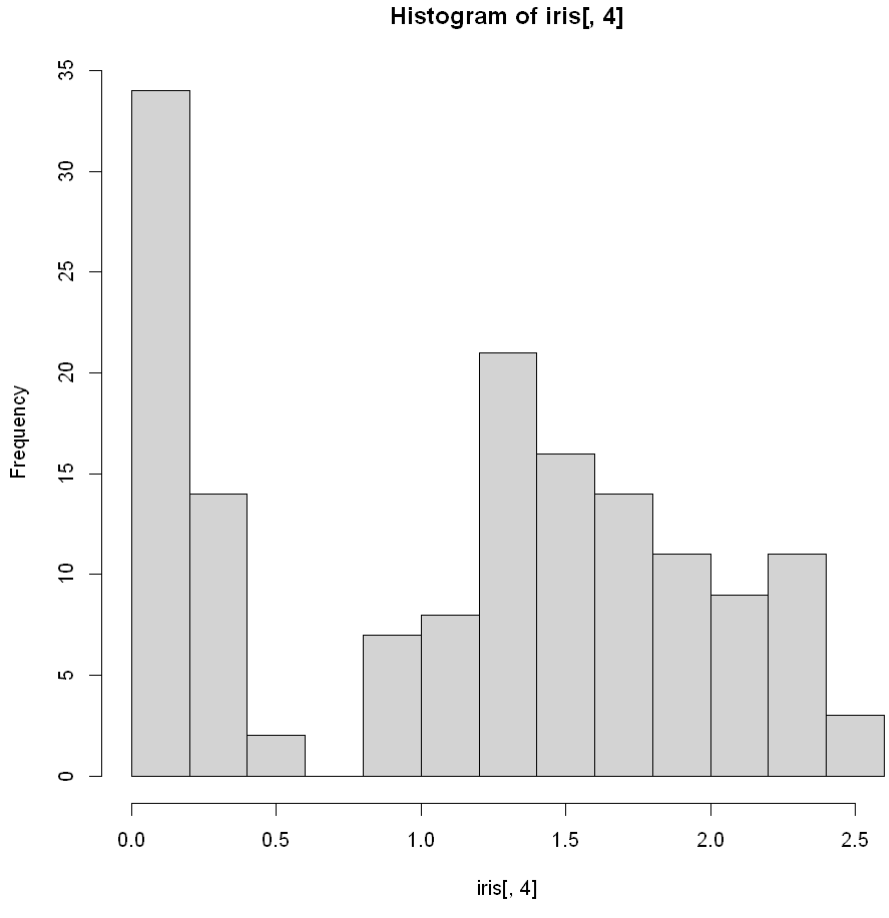
In [128]:
```
hist(iris[,2])
```



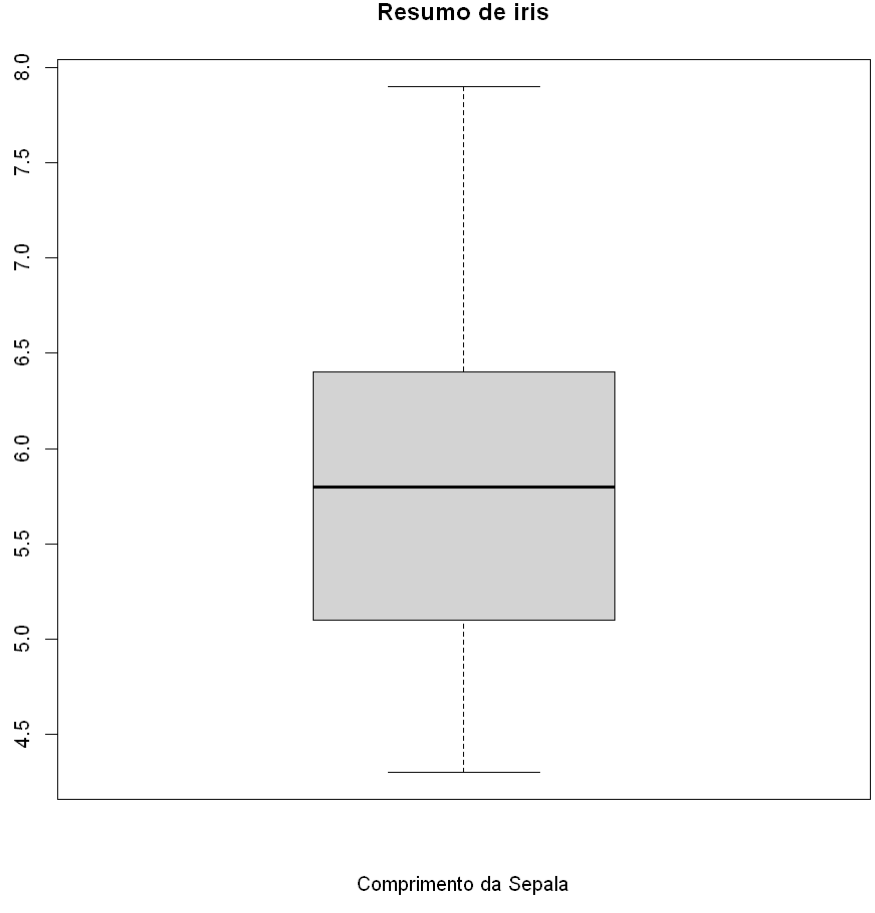
In [129]:
```
hist(iris[,3])
```



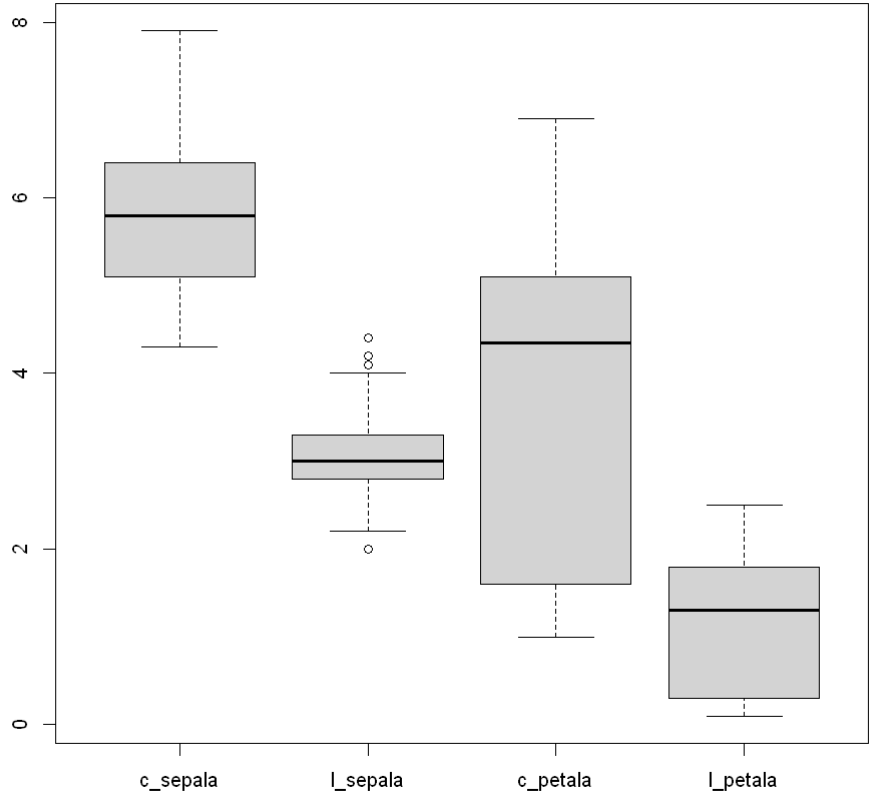
In [130]:
```
hist(iris[,4])
```



In [131]:
```
# O boxplot é usado para exibir o mesmo abaixo:
boxplot(iris$c_sepala,main="Resumo de iris",xlab="Comprimento da Sepala")
```



In [134]:
```
boxplot(iris[,1:4])
```



O valor da mediana é representado pela linha espessa na caixa e está entre 5,5 e 6,0 (5,75)
Também exibe os intervalos interquartis, bem como os valores mínimo e máximo.

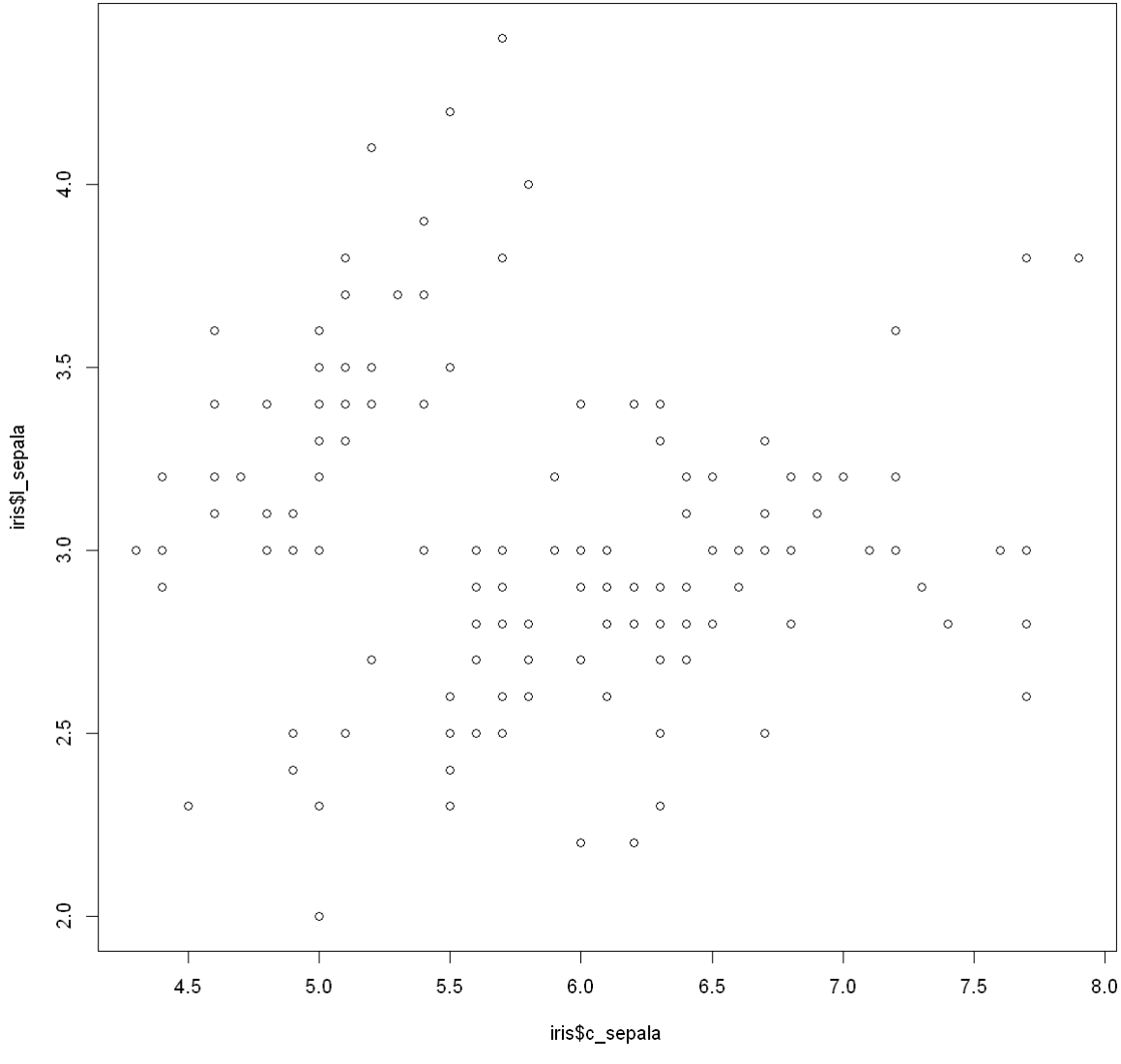
In [81]:
```
iris$c_sepala
```

5.1 · 4.9 · 4.7 · 4.6 · 5 · 5.4 · 4.6 · 5 · 4.4 · 4.9 · 5.4 · 4.8 · 4.8 · 4.3 ·
5.8 · 5.7 · 5.4 · 5.1 · 5.7 · 5.1 · 5.4 · 5.1 · 4.6 · 5.1 · 4.8 · 5 · 5 · 5.2 ·
5.2 · 4.7 · 4.8 · 5.4 · 5.2 · 5.5 · 4.9 · 5 · 5.5 · 4.9 · 4.4 · 5.1 · 5 · 4.5 ·
4.4 · 5 · 5.1 · 4.8 · 5.1 · 4.6 · 5.3 · 5 · 7 · 6.4 · 6.9 · 5.5 · 6.5 · 5.7 ·
6.3 · 4.9 · 6.6 · 5.2 · 5 · 5.9 · 6 · 6.1 · 5.6 · 6.7 · 5.6 · 5.8 · 6.2 · 5.6 ·
5.9 · 6.1 · 6.3 · 6.1 · 6.4 · 6.6 · 6.8 · 6.7 · 6 · 5.7 · 5.5 · 5.5 · 5.8 · 6 ·
5.4 · 6 · 6.7 · 6.3 · 5.6 · 5.5 · 5.5 · 6.1 · 5.8 · 5 · 5.6 · 5.7 · 5.7 · 6.2 ·
5.1 · 5.7 · 6.3 · 5.8 · 7.1 · 6.3 · 6.5 · 7.6 · 4.9 · 7.3 · 6.7 · 7.2 · 6.5 ·
6.4 · 6.8 · 5.7 · 5.8 · 6.4 · 6.5 · 7.7 · 7.7 · 6 · 6.9 · 5.6 · 7.7 · 6.3 · 6.7 ·
7.2 · 6.2 · 6.1 · 6.4 · 7.2 · 7.4 · 7.9 · 6.4 · 6.3 · 6.1 · 7.7 · 6.3 · 6.4 · 6 ·
6.9 · 6.7 · 6.9 · 5.8 · 6.8 · 6.7 · 6.7 · 6.3 · 6.5 · 6.2 · 5.9

In [135]:
```
summary(iris$c_sepala)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.300   5.100   5.800   5.843   6.400   7.900 
```

In [136]:
```
# EXPLORANDO A RELAÇÃO ENTRE O SEPAL COMPRIMENTO E A LARGURA USANDO PLOTS DE DISPERSÃO
options(repr.plot.width=10, repr.plot.height=10)
plot(iris$c_sepala, iris$l_sepala)
```



INSIGHTS DO SCATTER PLOT

1. O gráfico de dispersão sugere que pode não haver uma relação forte entre o comprimento e a largura da sépala.
2. A realização de modelos de regressão linear e análise de correlação pintará uma imagem melhor da relação.

A função de `plot` convencional faz o trabalho, mas para obter uma visualização melhor e mais aprimorada, Vamos usar o pacote `ggplot` para traçar o gráfico de dispersão

In [137]:
```
names(iris)
```

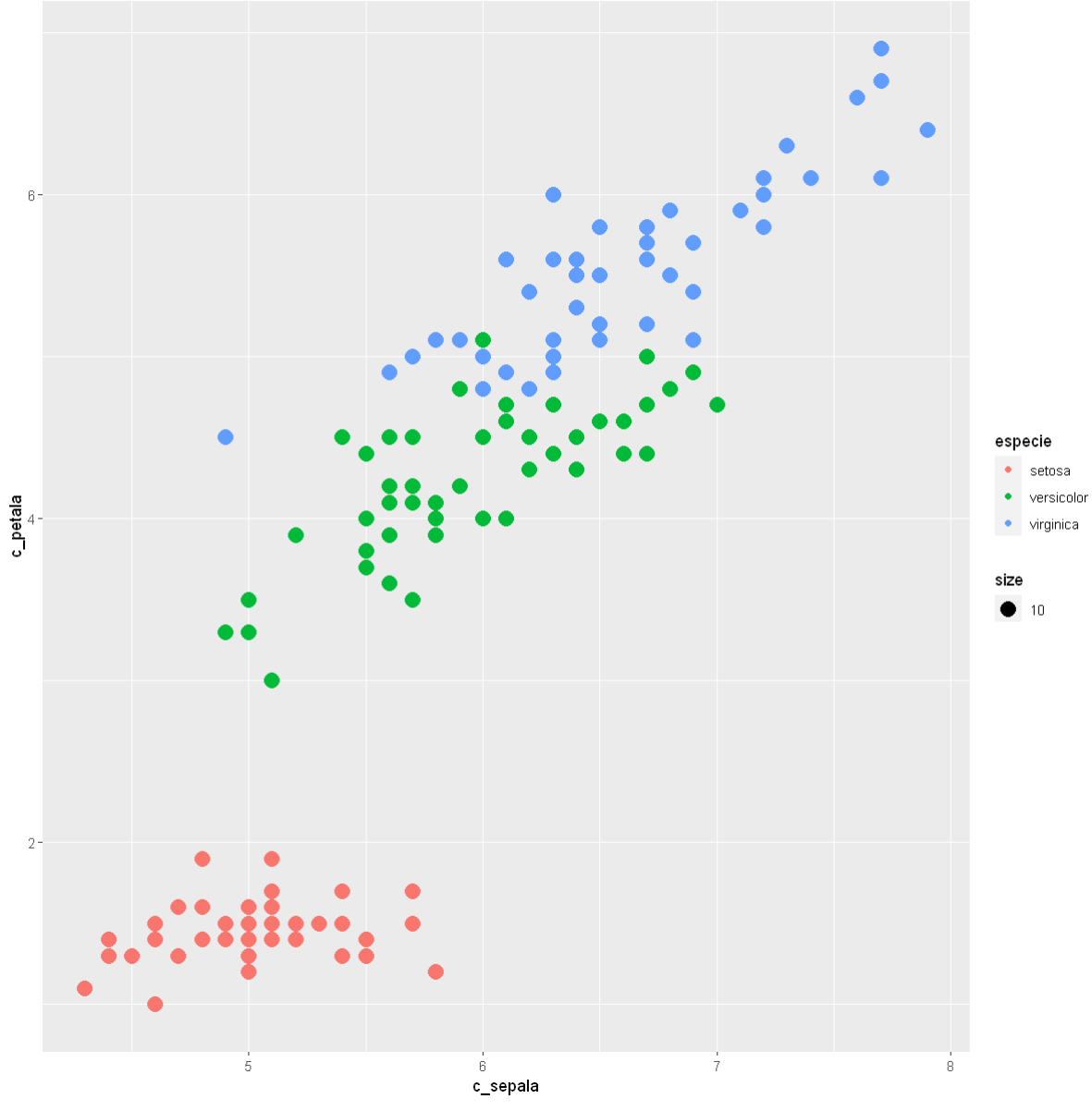
'c_sepala' · 'l_sepala' · 'c_petala' · 'l_petala' · 'especie'

In [138]:
```
install.packages("ggplot2")
```

```
Warning message:
"package 'ggplot2' is in use and will not be installed"
```

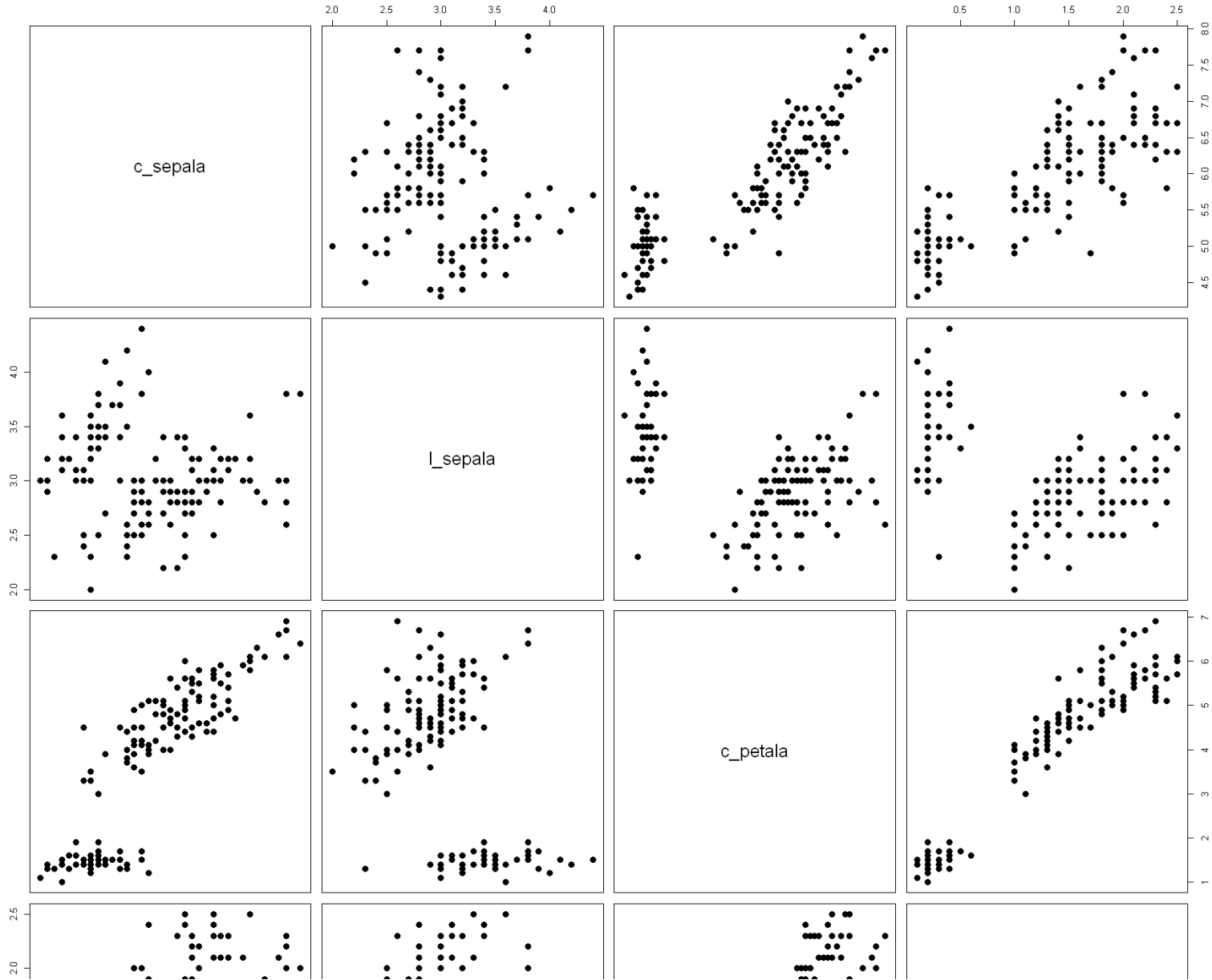
In [139]:

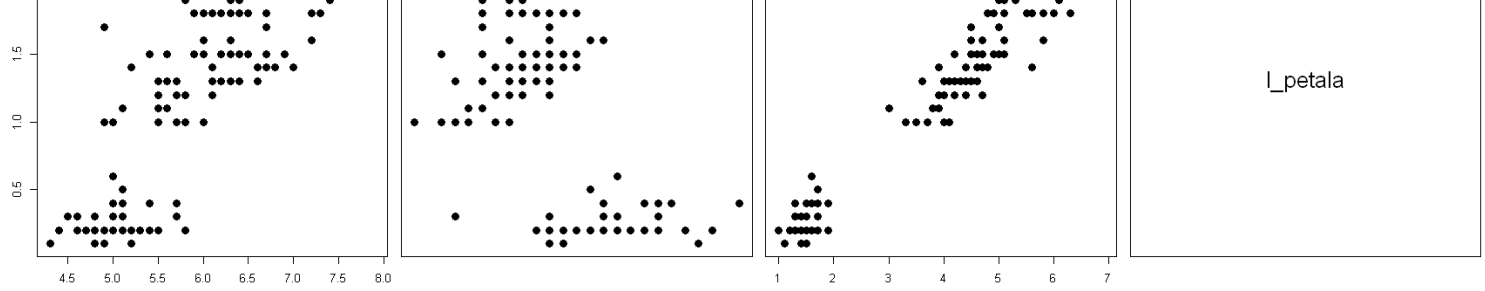
```
options(repr.plot.width=10, repr.plot.height=10)
library(ggplot2)
qplot(c_sepala, c_petala, data = iris, color = especie, size=10)
```



In [140]:

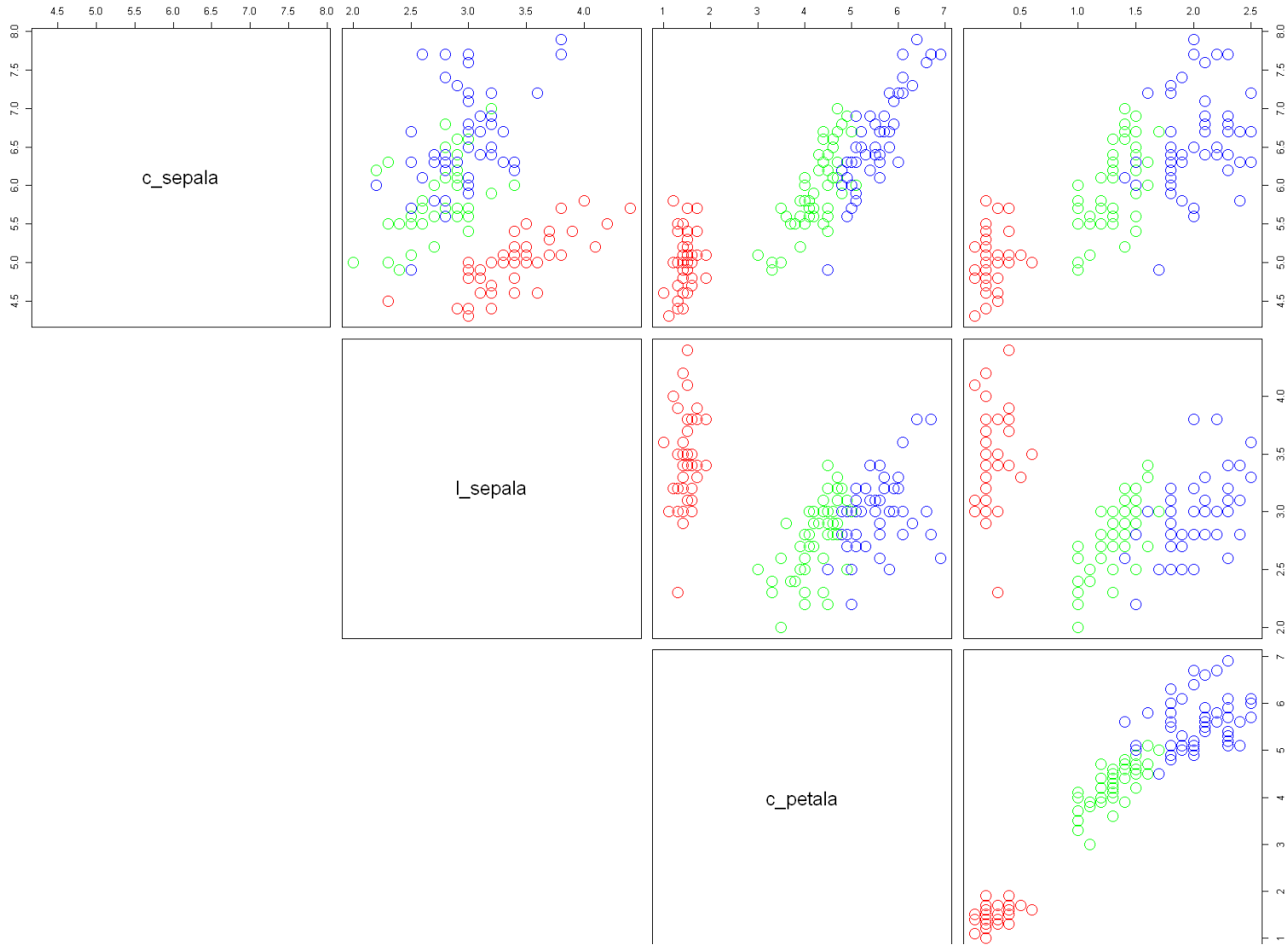
```
options(repr.plot.width=15, repr.plot.height=15)
pairs(iris[,1:4], pch = 19, cex=1.5)
```

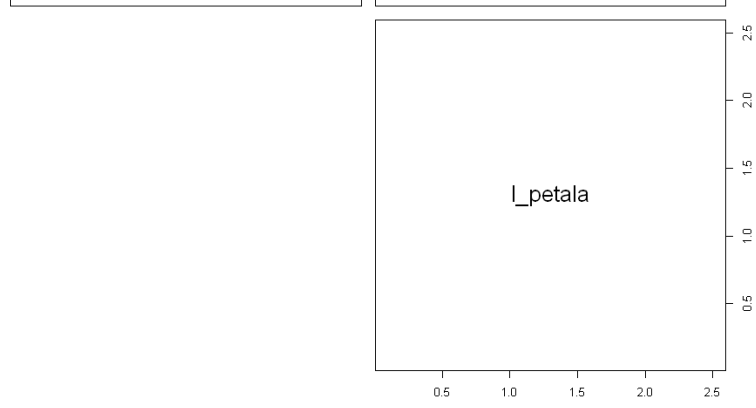



0.5
1.0
1.5
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
1
2
3
4
5
6
7
l_petala

In [141]:

```
my_cols <- c("red", "green", "blue")
options(repr.plot.width=15, repr.plot.height=15)
pairs(iris[,1:4], pch = 1,  cex = 2.5,
      col = my_cols[iris$especie],
      lower.panel=NULL)
```




l_petala
0.5
1.0
1.5
2.0
2.5