Análise Descritiva

Fabrício Jailson Barth

Abril de 2014

Acesso aos dados

```
require(UsingR)
data(survey)
head(survey)

data(iris)
head(iris)
```

Caracterização dos dados

No R, é possível testar se um atributo é **qualitativo** (factor) ou **quantitativo** (numeric).

```
is.numeric(survey$Pulse)
is.factor(survey$Sex)
is.numeric(survey$Smoke)
is.factor(survey$Height)

is.numeric(iris$sepallength)
is.factor(iris$class)

help(survey)
```

Caracterização dos dados

- A escala define as operacoes que podem ser realizadas sobre os valores do atributo.
- Em relação à escala, os atributos podem ser classificados como **nominais**, **ordinais**, **discreto** e **contínuo**.
- Os dois primeiros são do tipo qualitativo e os dois últimos são quantitativos.

- Na escala **nominal**, os valores são apenas **nomes diferentes**, carregando a menor quantidade de informação possível. Não existe uma relação de ordem entre seus valores.
- Os valores em uma escala **ordinal** refletem também uma ordem das categorias representadas. Dessa forma, além dos operadores de igualdade e desigualdade, operadores como $<$, $>$, $\geq$, $\leq$ podem ser utilizados.

- Baseado na descrição anterior, os atribudos dos datasets iris e jogar podem ser classificados como indicado nas tabelas abaixo:

```
survey$Pulse = quantitativo
survey$Sex = nominal (qualitativo)
survey$Smoke = ordinal (qualitativo)
survey$Height = quantitativo

iris$sepallength = racional (quantitativo contínuo)
iris$sepalwidth = racional (quantitativo contínuo)
iris$petallength = racional (quantitativo contínuo)
iris$petalwidth = racional (quantitativo contínuo)
iris$class = nominal (qualitativo)
```

Exploração de dados

Uma das formas mais simples de explorar um conjunto de dados é a extração de medidas de uma área da estatística denominada estatística descritiva. A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados. Tais características podem ser:

- Frequência;
- Localização ou tendência central (por exemplo, a média);
- Dispersão ou espalhamento (por exemplo, o desvio padrão);
- Distribuição ou formato.

No R é trivial identificar a média e mediana de um dado conjunto de valores para um atributo qualquer, como apresentado abaixo:

```
mean(survey$Pulse)
median(survey$Pulse)
```

Ou sumarizar todos estes valores através de um único comando:

```
summary(survey$Pulse)
```

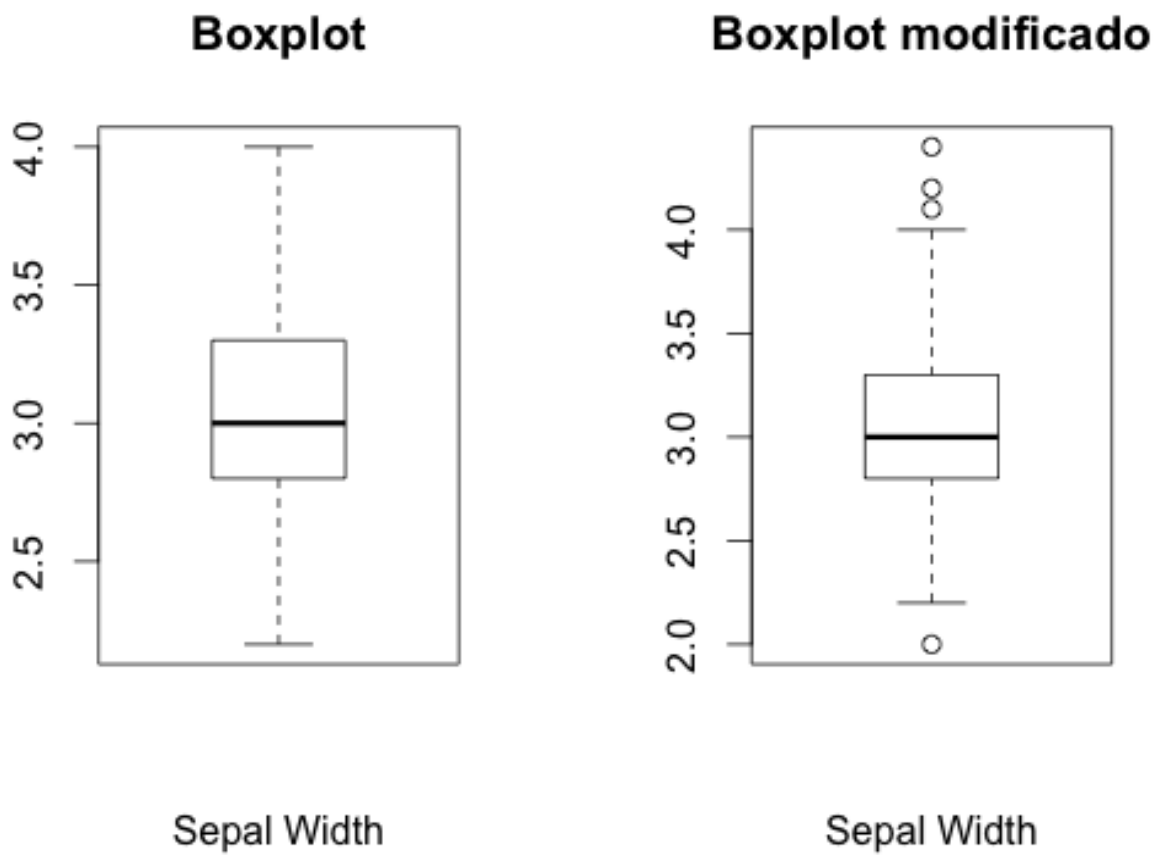
Além das informações textuais obtidas por

```
summary(iris$sepalwidth)
```

É possível obter um resumo visual da centralidade dos dados através do gráfico *boxplot*. No R é simples gerar este tipo de gráfico.

```
par(mfrow=c(1,2))
boxplot(iris$Sepal.Width,
        outline= FALSE, main="Boxplot",
        xlab="Sepal Width")
boxplot(iris$Sepal.Width,
        main="Boxplot modificado",
        xlab="Sepal Width")
```

Boxplot


Boxplot
4.0
3.5
3.0
2.5
Sepal Width
Boxplot modificado
4.0
3.5
3.0
2.5
2.0
Sepal Width

Boxplot modificado

O segundo gráfico ilustra uma variação do gráfico *boxplot*, conhecida como *boxplot* modificado. Neste gráfico, os valores acima do limite superior e abaixo do limite inferior são considerados *outliers*. Neste gráfico, 4 valores *outliers* são representados por círculos, 3 maiores que o 3o quartil + 1,5 x (3o quartil - 1o quartil) e 1 menor que 1o quartil - 1,5 x (3o quartil - 1o quartil).

Espalhamento de valores

As medidas mais utilizadas para avaliar o **espalhamento** de valores é a **variância** e o **desvio padrão**. Sendo que o desvio padrão é dado pela raiz quadrada da variância.

Desvio padrão:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (1)$$

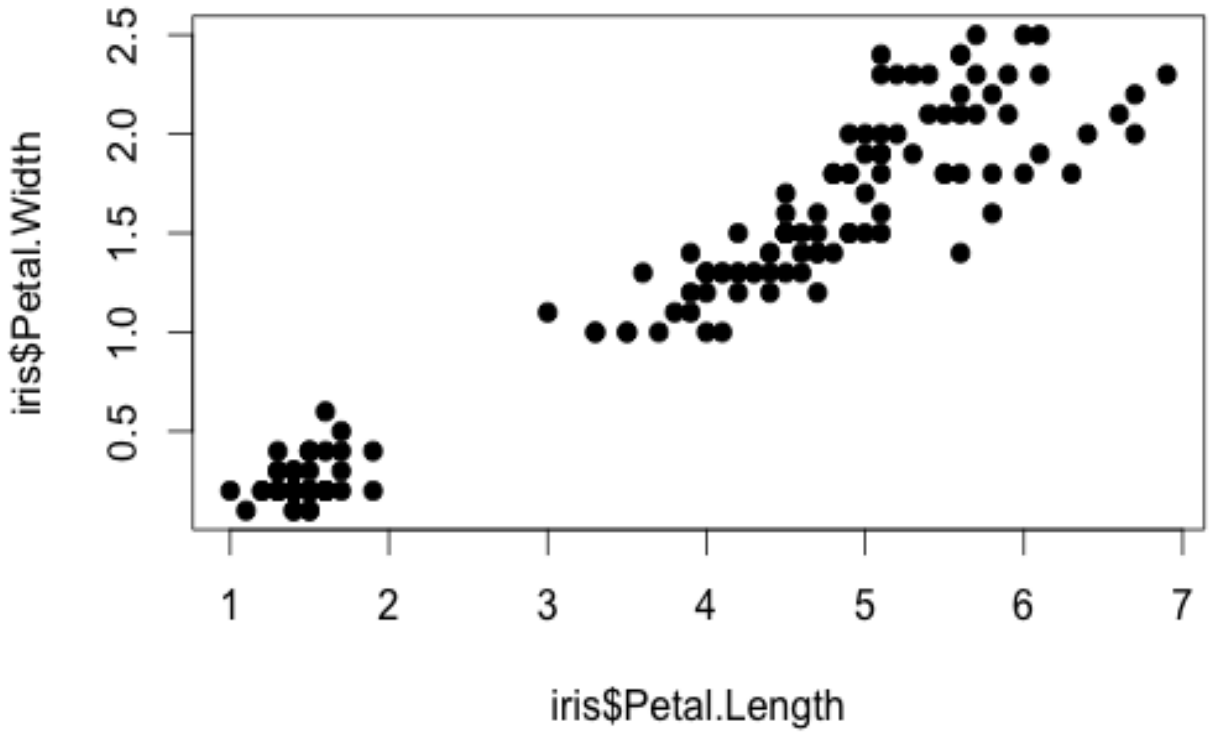
Variância:

$$s = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2}$$

```
var(iris$sepallength)
```

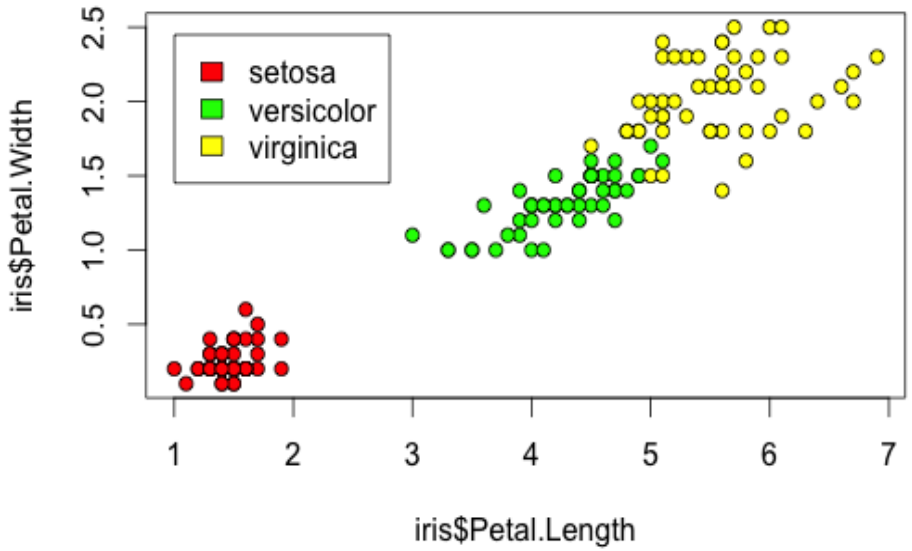
Plot

```
plot(iris$Petal.Length, iris$Petal.Width, pch=19)
```



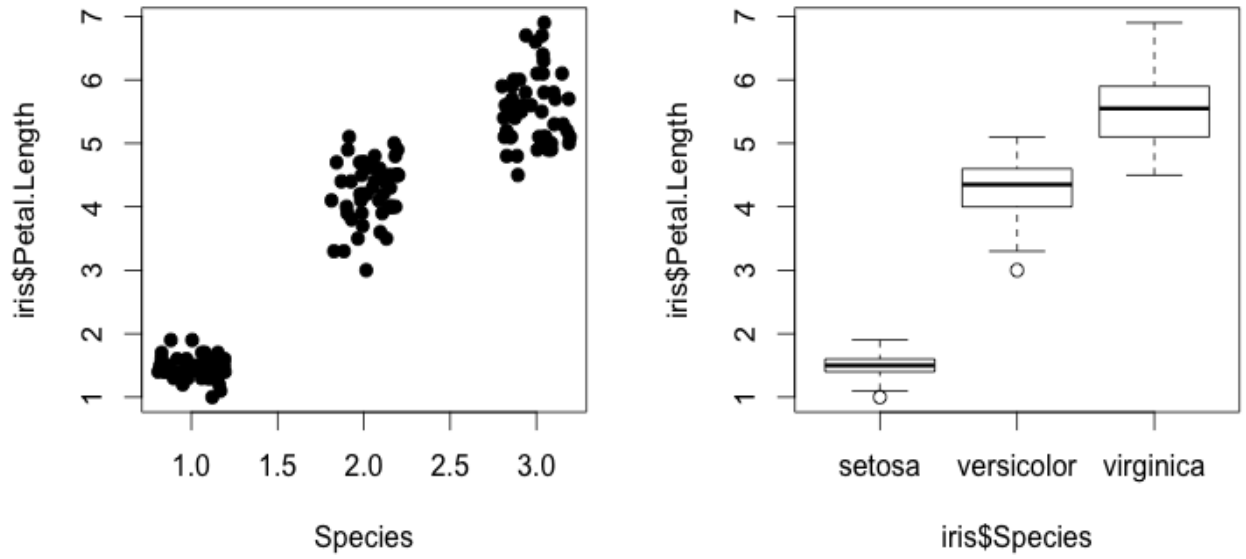
Plot

```
plot(iris$Petal.Length, iris$Petal.Width, pch=21,
     bg=c("red","green","yellow")[as.numeric(iris$Species)])
legend(locator(1), levels(iris$Species),
     fill=c("red","green","yellow"))
```



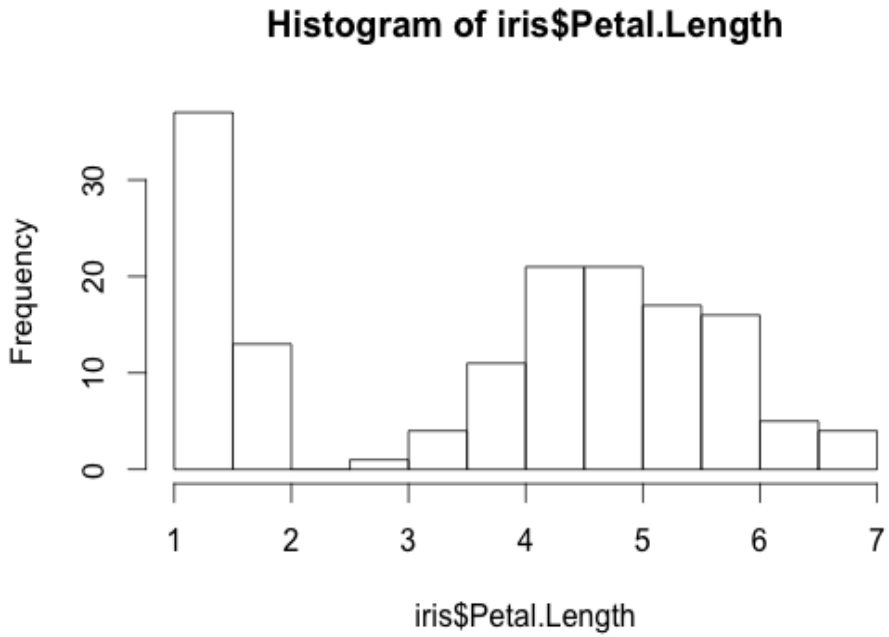
Comparando valores

```
par(mfrow=c(1,2))
plot(jitter(as.numeric(iris$Species)), iris$Petal.Length, pch=19, xlab="Species")
plot(iris$Petal.Length ~ iris$Species)
```



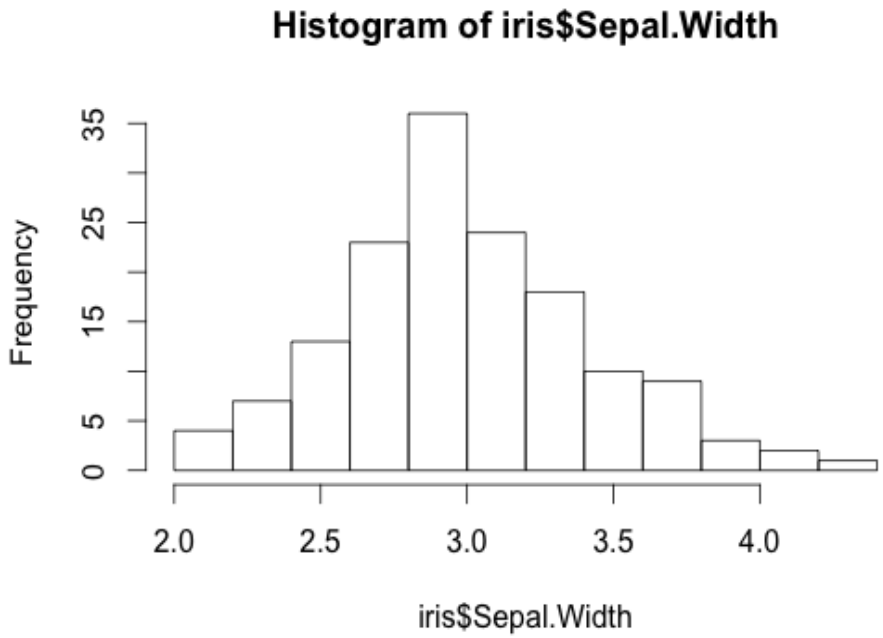
Histograma

```
> hist(iris$Petal.Length)
> summary(iris$Petal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.600   4.350   3.758   5.100   6.900
> var(iris$Petal.Length)
[1] 3.116278
```



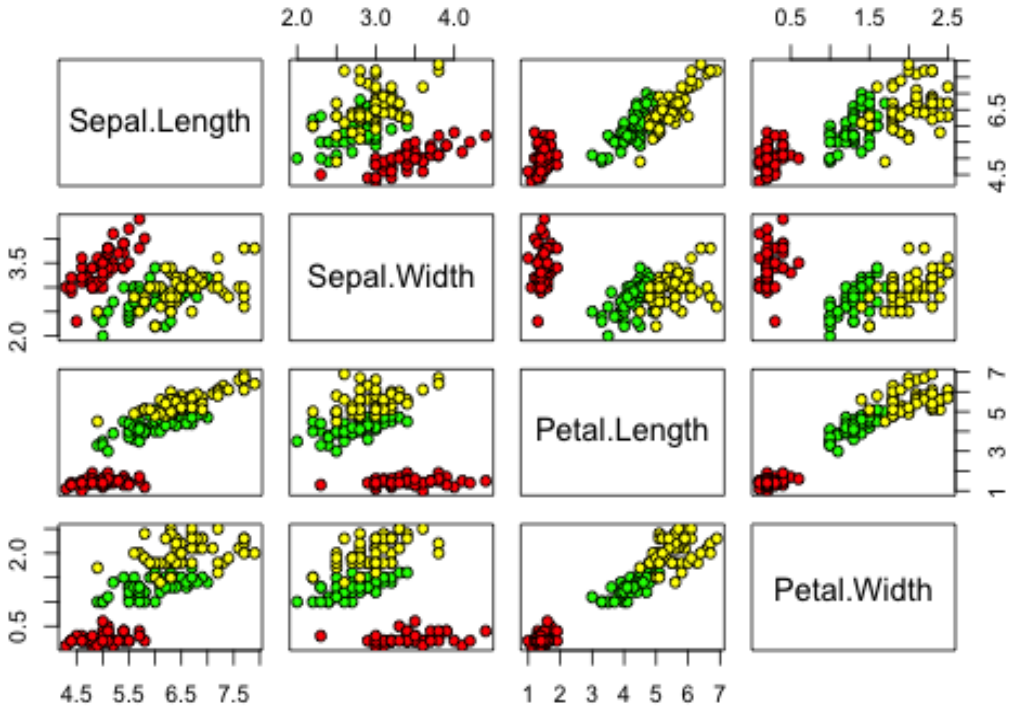
Histograma

```
> hist(iris$Sepal.Width)
> summary(iris$Sepal.Width)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.800   3.000   3.057   3.300   4.400
> var(iris$Sepal.Width)
[1] 0.1899794
```



Scatter Plot

```
plot(iris[,1:4], pch=21,
     bg=c("red","green","yellow")[as.numeric(iris$Species)])
```



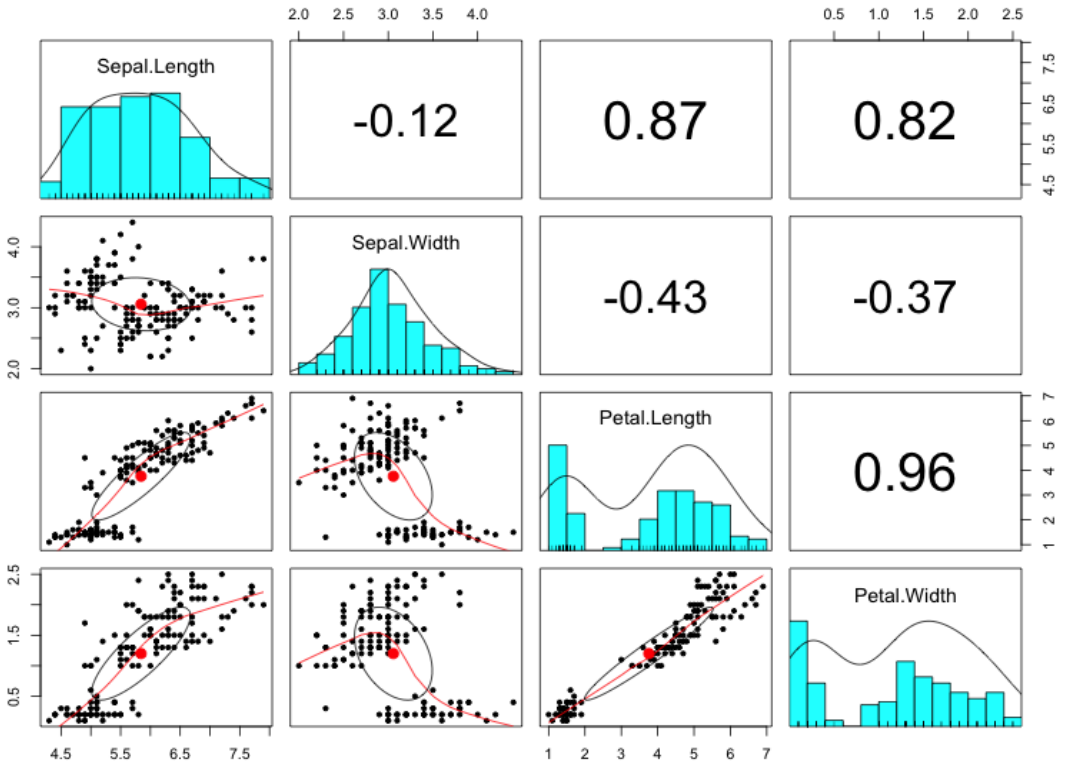
Correlação

- Dados multivariados permitem análises da relação entre dois ou mais atributos. Por exemplo, para atributos quantitativos, pode-se utilizar uma medida de correlação para identificar a relação linear entre dois atributos.

```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Resumindo a relação entre dados numéricos

```
library(psych)
pairs.panels(iris[,1:4])
```



Material de **consulta**

- Faceli, Lorena, Gama, Carvalho. Inteligência Artificial: uma abordagem de aprendizado de máquina, 2011. Capítulo 2: Análise de Dados.
- Cursos: *Computing for Data Analysis* e *Data Analysis* da www.coursera.org