ANÁLISE ESTATÍSTICA DE DADOS USANDO O RSTUDIO

Sheila Regina Oro[1]

8 de maio de 2018

[1] Universidade Tecnológica Federal do Paraná
Câmpus Francisco Beltrão
Departamento Acadêmico de Física, Estatística e Matemática
sheilaro@utfpr.edu.br

1 Apresentação

Este minicurso visa dar aos participantes uma introdução ao sistema estatístico R. Estudantes de graduação e pesquisadores formam o público alvo. O objetivo principal é apresentar o ambiente de trabalho *RStudio*, com ênfase na compreensão de princípios básicos da linguagem, da estrutura e a forma de operar o programa para a realização de análises estatísticas de conjuntos de dados. O objetivo específico é apresentar os comandos necessários para: iniciar, salvar e concluir uma sessão no R; instalar pacotes específicos para determinadas situações; importar arquivos de dados (.csv e .xls); obter estatísticas descritivas para uma ou mais variáveis; gerar gráficos exploratórios de análise de dados; aplicar testes de hipóteses para inferências sobre uma ou mais variáveis.

2 *Software* R e o ambiente RStudio

O software R é um programa livre, no sentido de possuir livre distribuição e código fonte aberto, compatível com as plataformas Windows, MAC e Linux. No endereço https://www.r-project.org , é possível obter o arquivo para a instalação do *software*.

A versatilidade é uma das qualidades do R, pois além dos pacotes disponibilizados em sua base, recebe contribuições de pesquisadores de todo o mundo na forma de novos pacotes e possibilita a criação de novas rotinas e funções. .

Para o usuário que possui pouca afinidade com linguagem de programação, o uso do ambiente RStudio pode facilitar a realização de atividades e melhorar a experiência com o *software*. O RStudio inclui um console, para a execução direta de comandos, bem como ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho. O RStudio é disponibilizado no endereço https://www.rstudio.com/products/rstudio/download , em versões gratuitas e comerciais,podendo ser executado na área de trabalho (Windows, Mac e Linux) ou em um navegador conectado ao RStudio Server ou RStudio Server Pro.

Dentre as muitas utilidades, destaca-se o uso do `R` para a análise estatística de dados. As seções a seguir apresentam os aspectos básicos para a utilização do ambiente RStudio e a sintaxe para a realização da análise estatística de dados.

3 Primeira Seção em R

O conteúdo apresentado no console de uma seção em **R** não pode ser salva, ou seja, os resultados não são conservados ao fechar o programa. Para contornar essa restrição basta criar um *script* (roteiro), inserir as linhas de comando (funções, requisição de pacotes, entrada de dados, comentários, por exemplo), e salvar para que sejam utilizadas novamente quando necessário.

Para o funcionamento correto e evitar o surgimento de erros de execução dos comandos no **R**, alguns cuidados simples devem ser tomados, pois esquecimento um parênteses, uma vírgula ou um ponto, por exemplo, pode gerar erros e impedir a obtenção dos resultados almejados.

Para a criação de um *script*, primeiramente clique em "File → New script". Automaticamente abrirá uma janela, em seguida, clique em "File → Save as...", para salvar no local desejado o *script*.

Para executar as ações das linhas de comandos digitadas no *script*, basta posicionar o cursor na respectiva, ou selecionar as linhas de comandos que deseja executar e pressionar as teclas `Ctrl + R`.

Algumas observações importantes:

- O símbolo `#` é utilizado para inserir um comentário no *script*;
- Os símbolos `<-` e `=` são utilizados para atribuir nomes às funções, ou comados;
- Para facilitar a identificação, os comandos em **R** estarão destacados pela cor azul no texto das seções a seguir, ou seja, são comandos escritos exatamente na forma que o

software fará a leitura. As observações inseridas após o símbolo `#`, ou entre estes, são comentários que servem para explicar ou destacar determinado comando.

- Para visualizar todos os argumentos que uma função comporta e explicações sobre eles, basta colocar o símbolo `?` antes do comando, ou então digitar `help()`. Por exemplo `?seq` ou `help(seq)`.

- As operações básicas de adição, subtração, multiplicação e divisão podem ser executadas no software `R` utilizando os operadores "+, -, *, /" respectivamente. A seguir são apresentados alguns exemplos de operações básicas.

Pratique!

Exercício 1: Crie um *script* e salve-o no local que desejar com o nome MINICURSO. Escreva no início o seguinte comentário:

"Roteiro para a realização de análises estatísticas de dados.
Data: (digite a data de hoje).
Autor: (insira seu nome)."

3.1 Entrada dos dados

- Vetor com elementos pré-definidos (números ou caracteres)

```
x <- c(1,0,-2,3,4,6,10)
fruta <- c("banana", "abacaxi", "morango", "uva")
```

- Seleção de elementos de um vetor, com posição definida

```
x[4]
# Seleciona o 4º elemento do vetor x

x[1:4]
# Seleciona do 1º ao 4º elementos do vetor x
```

- Inserção de uma matriz

```
M <- matrix(c(2,4,6,8,10,12), nrow = 3, ncol = 2)
# Armazena em M uma matriz com 3 linhas e 2 colunas

M[,2]
```

```
# Seleciona os elementos da 2ª coluna da matrix M

M[3,]
# Seleciona os elementos da 3ª linha da matrix M

M[1,3]
# Seleciona o elemento da 1ª linha e 3ª coluna da matrix M
```

- Banco de dados disponível no R

```
data(iris)
# Carrega o banco de dados "iris"

A <- iris[1:25,]
# Armazena em A as primeiras 25 linhas do banco de dados "iris"

names(A)
# Mostra o nome das variáveis (colunas) armazenadas em A

attach(A)
# Guarda o nome das variáveis de A

A$Petal.Length
# Escolha de uma variável (Petal.Length) armazenada em A

iris[iris$Sepal.Length > 6,]
# Seleção de amostra com tamanho de sépala maior que 6

iris[iris$Species == "virginica",]
# Seleção de amostra da espécie "virginica" ##
```

- Importação de dados

  - Planilha do Excel (.xls): "Import DataSet"→ "From Excel"→ "Browse"→ (selecione o arquivo no local em que ele está armazenado) → "Import". Também é possível importar dados utilizando os comandos

    ```
    library(readxl)
    dados <- read_excel("~/local/nome_da_planilha.xls")
    ```

– Arquivo de texto (.csv): "Import DataSet"→ "From CSV"→ "Browse"→ (selecione o arquivo no local em que ele está armazenado) → "Import". Também é possível importar dados utilizando os comandos:

```
library(readr)
dados <- read_csv("~/local/nome_do_documento.csv")
```

Pratique!

Exercício 2: Armazene os valores 5, 4, 6, 8, 12 num vetor x. Selecione o terceiro elemento do vetor.

Exercício 3: Escreva a matriz

$$M = \begin{bmatrix} 2 & 2 & 3 \\ 1 & -1 & 5 \\ 4 & 0 & -2 \\ 0 & 7 & -1 \end{bmatrix}$$

Exercício 4: Crie uma planilha no Excel com os valores da matriz M do exercício anterior. Nomeie as colunas da planilha como C1, C2, C3.

a) Salve e importe a planilha com o nome "amostra.xls".

b) Salve e importe a planilha com o nome "valores.csv".

4 Análise Exploratória dos Dados

4.1 Medidas descritivas

Para obter os valores das principais medidas de posição: Mínimo, 1º Quartil, Mediana, Média, 3º Quartil e Máximo, usa-se o comando *summary*.

```
summary(dados)
```

Para obter as medidas de posição de grupos específicos de um conjunto de dados, usa-se o comando *tapply*.

```
tapply(Sepal.Length,Species, summary)
```

As medidas de dispersão são obtidas por meio de comandos específicos:

```
var(dados) # Variância
sd(dados) # Desvio padrão
cv <- 100*(sd/mean(A)) # Coeficiente de variação
```

4.2 Gráficos e Diagramas

```
boxplot(dados) # Gráfico Boxplot
boxplot (resposta~fator)  # Gráfico Boxplot estratificado
hist(dados) # Histograma
stem(dados) # Diagrama Ramo-folhas
```

Pratique!

Exercício 5: Habilite o banco de dados "ChickWeight". Os dados referem-se ao peso dos frangos (weight), de acordo com a dieta (Diet) e o tempo (Time).

a) Obtenha as medidas descritivas para o peso dos frangos (geral).

b) Obtenha as medidas descritivas para o peso dos frangos de acordo com a dieta.

c) Obtenha as medidas descritivas para o peso dos frangos de acordo com o tempo.

d) Gere os gráficos Boxplot (geral), Boxplot estratificado de acordo com a dieta e Boxplot estratificado de acordo com o tempo.

e) Gere o Diagrama Ramo-folhas para o peso dos frangos.

5 Testes de Hipóteses

5.1 Testes de Normalidade

O valor-p deve ser superior a 5% para aceitar a hipótese de normalidade.

```
library(nortest) # Carrega o pacote "nortest"
qqnorm(resposta) # Gráfico QQPlot
qqline(resposta) # Insere a reta no QQplot
ad.test(resposta) # Teste de Anderson-darling
shapiro.test(resposta) # Teste de Shapiro-Wilk
lillie.test(resposta) # Teste de Kolmogorov-Smirnov
```

5.2 Teste de Homogeneidade

O valor-p deve ser superior a 5% para os grupos serem considerados homogêneos.

```
library(car) # Carrega o pacote "car"
leveneTest(resposta~tratamento)
```

5.3 Transformação BOX-COX

Usada, em geral, quando as suposições de normalidade e/ou homogeneidade forem violadas.

```
require(MASS)
lambda <- boxcox(resposta ~ tratamento, plotit=T, lam=seq(-1, 1, 1/10))
```

Se ficar difícil de visualizar, aplicar zoom usando o comando lambda = seq()

```
lambda <- boxcox(resposta ~ tratamento, plotit=T, lambda = seq(-0.15, 0.65, len = 20))
lambda
```

Realizar a transformação com o valor ótimo identificado

```
 resposta transformada = (resposta^(lambda otimo) - 1)/ lambda otimo
respostat <- (resposta^(0.3) - 1)/0.3

install.packages("multcomp", dep=TRUE)
```

5.4 Teste t

Usado para comparar as médias de dois tratamentos.

```
# Exemplo de entrada dos dados
Lines <-
"fator resposta
A 0.713
A 0.635
A 0.757
A 0.621
A 0.527
B 0.734
B 0.635
B 0.763
B 0.597
B 0.415
B 0.460"


variavel <- read.table(textConnection(Lines), header=TRUE);
```

```
closeAllConnections()
str(variavel)

t.test(resposta~fator,data=variavel,var.equal=TRUE, alternative="two.sided")
```

5.5 ANOVA - 1 Fator

Compara a variação devida aos tratamentos com a variação devido ao acaso.

```
anova <-aov(resposta~fator, data=dados)
summary(anova) # Tabela da ANOVA
coef(anova) # Coeficientes do modelo
residuals(anova) # Valores dos Resíduos do modelo
fitted(anova) # Valores ajustados
plot(residuals(anova), type="l") # Gráfico dos resíduos (para verificar a independência)
```

5.6 Teste de Tukey

Usado para comparar as médias de múltiplos tratamentos, quando a ANOVA apontar para diferença significativa entre os eles.

```
TukeyHSD(anova, "fator", ordered = TRUE)
plot(TukeyHSD(anova, "fator"))
```

Pratique!

Exercício 6: Considere os dados referentes à Salinidade de três Lagunas.

a) Faça o teste da ANOVA.

b) Aplique o teste de Tukey.

c) Verifique a Normalidade dos Resíduos (do modelo ANOVA).

Exercício 7: Compare a produção de batatas (em toneladas) de acordo com a variedade.

```
batata <- scan ()
9.2
13.4
11
9.2
21.1
27
26.4
```

```
25.7
22.6
29.9
24.2
25.1
15.4
11.9
10.1
12.3
12.7
18
18.2
17.1
20
21.1
20
28
23.1
24.2
26.4
16.3
18
24.6
24
24.6

producao <- data.frame(variedade = factor(rep(1:2, each=16)), resp=batata)
```