



UFFS – Universidade Federal da Fronteira Sul
Campus: Cerro Largo
Professora: Iara Denise Endruweit Battisti
e-mail: lara.battisti@uffs.edu.br

Oficina Semana Acadêmica do curso de Física – licenciatura

Introdução ao Software R

1 Introdução

1.1 Download

Link para download do software R e RStudio:

www.r-project.org

www.rstudio.com/products/rstudio/download

Instalação do R e RStudio através dos arquivos executáveis

R-3.301-win.exe

RStudio-0.99.902.exe

R é o software

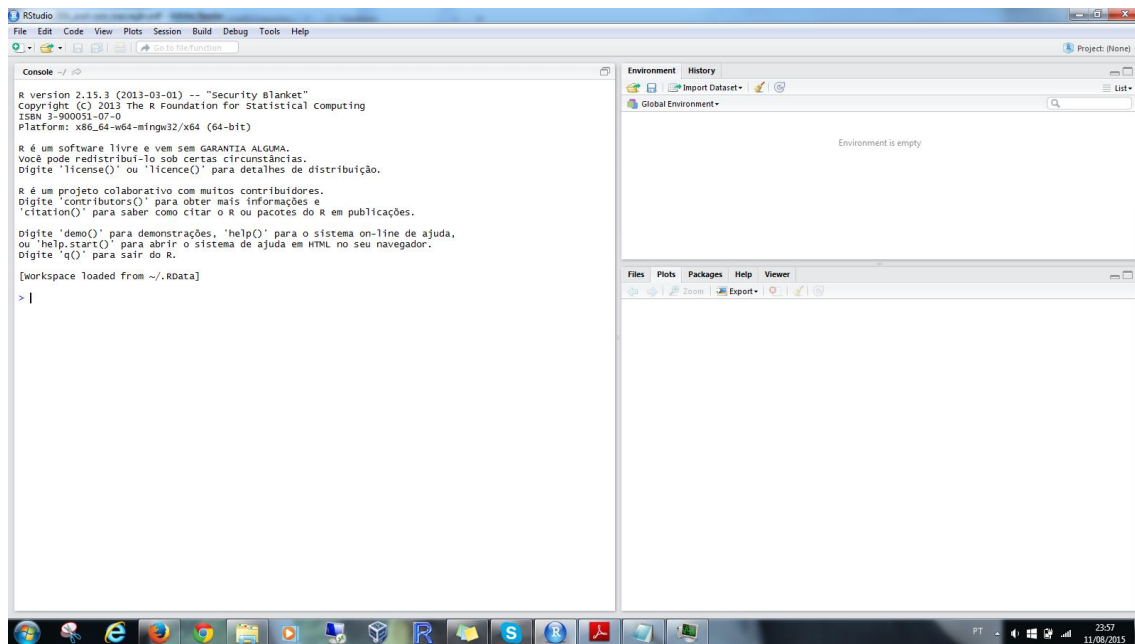
Rstudio é uma interface amigável para o R

Curso de extensão do Software R – UFFS, campus Cerro Largo

<https://smolski.github.io/softwarelivrer/index.html>

1.2 Paineis do RStudio

Conhecendo cada painel (janela) do RStudio:

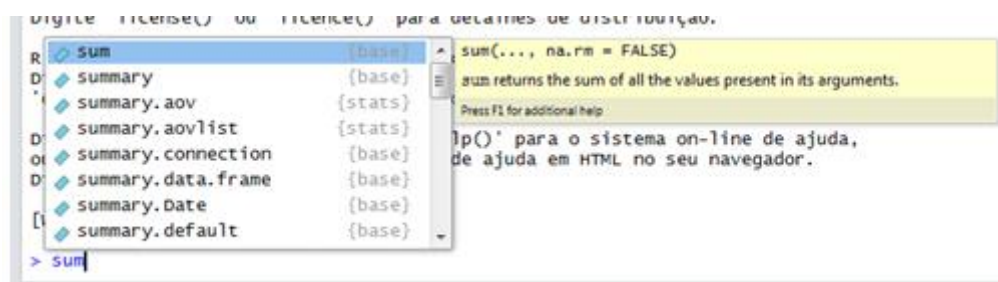


Digitando um comando na linha de comando (painel console)

O caracter “>” é apresentado no início da linha de comando aguardando um comando.

Enquanto estamos digitando um comando, abre uma janela com opções para esse comando, por exemplo:

- para o comando summary:



E logo depois aparece uma outra janela (amarela) com informações sobre a função, como: o que faz, quais seus argumentos, etc.

Se a janela de sugestões não aparecer, basta aperta CTRL+Espaço.

Para limpar o console: CTRL+L ou menu Edit e clicar em 'clear console'.

Criando um Script (painel de script)

Clicar no ícone com um + no canto esquerdo da barra de ferramentas ou no menu 'File -> New File -> R Script'. Uma nova janela acima da janela do console será aberta.

Podemos executar diretamente pela interface gráfica, sem precisar carregá-lo no comando de linha. Por exemplo, se dermos dois comandos simples no script:

```
>amostra = rnorm(10, 5, 3)
```

```
>mediaAmostrai = mean(amostra)
```

Após clicar no botão 'Source' no canto superior direito da janela do script, então todos os comandos serão executados.

Para esse exemplo, podemos ver no painel direito superior: as variáveis 'amostra' e 'mediaAmostrai' com seus valores.

Painel direito superior

No painel direito superior, a aba 'Environment' tem alguns elementos interessantes. Por exemplo:

- o ícone do disquete permite salvar as variáveis do espaço de trabalho;
- e a pastinha permite carregá-la.

O ícone 'Import dataset' permite importar dados em fomato texto (CSV, por exemplo).

A vassoura limpa o espaço de trabalho.

Na aba 'History' aparece a lista dos comandos já digitados.

Também, podemos fazer uma busca no histórico. Digitando o comando (palavra) de interesse na caixinha de texto com uma lupa. Assim, o RStudio encontra e mostra os comandos que tenham as letras digitadas. É uma boa maneira de recuperar um comando utilizado no comando de linha para passar para um script (utilizando o ícone "to source"), como por exemplo, para lembrar qual comando foi utilizado para gerar um gráfico.

Painel direito inferior

A aba 'Files' é utilizada para navegar nos dados do computador. Se clicar num arquivo compatível com o RStudio, ele abre imediatamente.

Na aba 'Plots' são mostrados os gráficos. Por exemplo, se digitarmos um comando para gerar um histograma no comando de linha: `hist(rnorm(50, 10, 5), border='white', col='skyblue', main='Histograma')`. Então, o RStudio irá gerar o gráfico.

É possível salvar o gráfico na aba 'Export'.

Na janela 'Plots' tem duas flechas (esquerda e direita) para navegar no histórico de gráficos.

Por exemplo: digitar os seguintes comandos no arquivo de script e depois cliquem em 'Source':

```
>X <- rnorm(50, 10, 5)
>Y <- 10 + 5*X + rnorm(50, 0, 10)
>plot(X, Y, pch=19, col='skyblue')
```

Então é gerado um gráfico de dispersão.

Assim, não é necessário salvar um gráfico imediatamente. É possível ir e voltar entre os gráficos criados durante uma sessão. O botão 'Zoom' abre o gráfico numa nova janela, um pouco maior.

O botão com um 'X' serve para apagar o gráfico atual e a vassoura limpa todo o histórico de gráficos.

A aba 'Packages', facilita carregar e instalar novos pacotes para o R.

Help

O help é acessado clicando no nome de um pacote ou usando o comando '`help()`' ou buscando direto dentro da aba 'Help'.

Por exemplo, se digitarmos:

```
help(plot)
```

O RStudio abrirá a aba help com o arquivo de ajuda do comando 'plot'.

1.3 Abrir arquivo de dados

Primeiramente vamos criar um banco de dados na planilha eletrônica LibreOffice Calc (ou EXCEL).

Vamos utilizar como exemplo o banco de dados apresentado a seguir, com o nome do arquivo 'arvores':

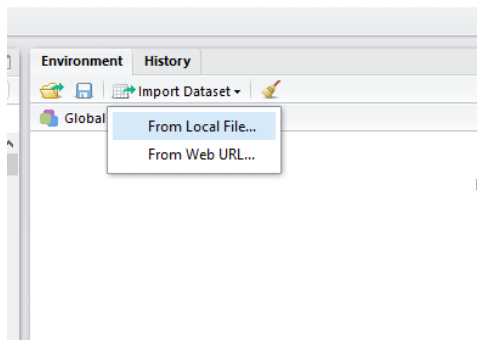
Os dados provem de uma pesquisa com espécies de árvores registrando as variáveis diâmetro altura do peito (DAP) e altura. Dados cedidos pela professora Tatiane Chassot.

Ler os dados pela linha de comando:

```
> arvores <- read.csv("D:/arvores.csv", sep=';', header = T)
> view(arvores)
```

Ou

Ler os dados no 'Environment' pelo 'Import Dataset...From Local File', conforme abaixo:



Após clicar em 'From Local File' é apresentada a seguinte janela:

Import Dataset

Name:

Encoding:

Heading: ☒ Yes ☐ No

Row names:

Separator:

Decimal:

Quote:

Comment:

na.strings:

☒ Strings as factors

Input File

```
Nomecientifico;diametro_cm;altura_m
Alsophila sp.;11,7;5,1
Alsophila sp.;11,3;6,0
Alsophila sp.;11,0;5,2
Alsophila sp.;14,0;3,5
Alsophila sp.;12,1;5,8
Araucaria angustifolia;62,1;24,2
Araucaria angustifolia;28,0;20,3
Araucaria angustifolia;35,3;18,1
Araucaria angustifolia;81,8;22,8
Araucaria angustifolia;10,2;9,4
Araucaria angustifolia;12,1;13,6
Araucaria angustifolia;16,7;14,1
Araucaria angustifolia;9,7;9,7
Araucaria angustifolia;12,7;15,2
Araucaria angustifolia;13,0;11,1
Araucaria angustifolia;11,6;15,3
Araucaria angustifolia;10,7;10,9
Araucaria angustifolia;21,2;18,9
```

Data Frame

Nomecientifico	diametro_cm	altura_m
Alsophila sp.	11,7	5,1
Alsophila sp.	11,3	6,0
Alsophila sp.	11,0	5,2
Alsophila sp.	14,0	3,5
Alsophila sp.	12,1	5,8
Araucaria angustifolia	62,1	24,2
Araucaria angustifolia	28,0	20,3
Araucaria angustifolia	35,3	18,1
Araucaria angustifolia	81,8	22,8
Araucaria angustifolia	10,2	9,4
Araucaria angustifolia	12,1	13,6
Araucaria angustifolia	16,7	14,1
Araucaria angustifolia	9,7	9,7
Araucaria angustifolia	12,7	15,2
Araucaria angustifolia	13,0	11,1
Araucaria angustifolia	11,6	15,3
Araucaria angustifolia	10,7	10,9
Araucaria angustifolia	21,2	18,9

Para entender o comando:

- read.csv é porque o tipo de arquivo é csv (arquivo texto separado por vírgula);
- sep=';' é para identificar o caracter separador;
- 'header' é para identificar a primeira linha como nome de variável.

Importante, o 'Open file' abre o arquivo para edição, mas não carrega os dados na memória como banco de dados, então é necessário utilizar outro comando antes de executar as estatísticas, o qual será apresentado mais adiante.

Desta forma, o banco de dados é apresentado em uma nova janela no R:

teste.Rmd x arvores x			
Filter			
	Nomecientifico	diametro_cm	altura_m
1	Alsophila sp.	11.7	5.1
2	Alsophila sp.	11.3	6.0
3	Alsophila sp.	11.0	5.2
4	Alsophila sp.	14.0	3.5
5	Alsophila sp.	12.1	5.8
6	Araucaria angustifolia	62.1	24.2
7	Araucaria angustifolia	28.0	20.3
8	Araucaria angustifolia	35.3	18.1
9	Araucaria angustifolia	81.8	22.8
10	Araucaria angustifolia	10.2	9.4

Showing 1 to 10 of 679 entries

1.4 Salvar arquivo de dados

O banco de dados que o R armazena na memória pode ser salvo, junto com todo o ambiente, usando o ícone de disquete na aba 'Environment', assim o arquivo é salvo com a extensão 'RData'. Para carregar o arquivo de dados, clicamos no ícone de pasta (Abrir dados...) na mesma aba.

Ou

Outra opção com mesmo efeito é utilizar o comando `save('nomeDoObjeto', file='nomeDoArquivo.RData')`. O nome do objeto pode ser uma lista de objetos para salvar mais de um objeto do ambiente, `'list=('objeto1', 'objeto2')`. Para carregar um arquivo RData no ambiente, o comando é `load('arquivo.RData')`, desde que o arquivo esteja no diretório de trabalho do R.

Ou

Outra opção é salvar só o objeto que contém o banco de dados como arquivo com extensão CSV, com o `write.csv ('nomeDoObjeto', file='nomeDoArquivo.RData')`. O objeto, porém, precisa ser uma matriz ou data frame, ou pelo menos ser convertível para esse tipo de objeto.

Podemos salvar o histórico dos comandos, utilizando o ícone de salvar na aba History. E quando desejar abrir o histórico na mesma aba.

1.5 Objetos

Atribuição: `'x = y'` é equivalente a `'x <- y'`.

Assim, uma variável com nome `'x'` é criada e dada a ela o valor da variável `'y'`.

Por exemplo, podemos criar uma variável `'arvores'` usando o comando `'read.csv'`:

```
arvores <- read.csv("D:/arvores.csv", sep=';', header = T)
```

1.6 Matriz

Para acessar a coluna cujo nome é `'curso'` podemos usar o comando `arvores$altura_m`.

Caso não lembrarmos os nomes das variáveis, podemos usar o comando: `names(arvores)` ou `colnames(arvores)`.

Para alterar o nome de uma variável podemos usar comando `'colnames'` e a notação para seleção de elemento de vetor: `colnames(arvores)[2] = 'DAP'`. Se necessário, para atualizar no janela do banco de dados, clica sobre o nome da variável com o botão direito do mouse e seleciona `'reload'`.

Então, com os nomes das variáveis, podemos acessá-las diretamente usando o \$, como por exemplo: `arvores$Nomecientifico`.

Ou então podemos usar a notação de matriz do R: `arvores[1]` ou `arvores[,1]`

O interessante da matriz é que podemos selecionar várias variáveis ao mesmo tempo. Por exemplo: `arvores[1:200,1:2]`

Para entender o comando:

- o nome da variável que armazena o banco de dados;
- abre colchetes para indicar que queremos 'recortar' a matriz de dados;
- depois do colchete, um número indicando que linha queremos recortar. Se não colocamos nenhum número, o R entende que é para todas as linhas. No exemplo, '1:200' indica que queremos os registros 1 a 200;
- depois vírgula;
- depois da vírgula, a coluna que desejamos recortar. Novamente, se deixamos em branco, o R entende que são todas as colunas. No exemplo, '1:2' indica que queremos as variáveis (colunas) 1 a 2;

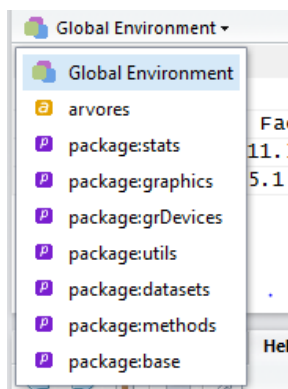
Também, podemos excluir observações ou variáveis, usando números negativos. Por exemplo, se queremos todas as variáveis, menos as duas últimas, basta acrescentar um '-': `arvores[-(2:3)]`

1.7 Linha de comandos



Porém não precisamos acessar as variáveis usando \$ ou [,]. Podemos usar o comando seguinte para acessar os dados e variáveis:

```
>attach(arvores)
```

Depois, na aba Environment (canto superior direito do RStudio) clica no botão 'Global environment' e seleciona o nome 'dados' que é o espaço dessa variável. Assim, aparecerá a lista de variáveis que foram anexadas ao espaço de trabalho a partir do banco 'dados'.



Resultado:

Environment		History
 Import Dataset		
arvores		
values		
altura_m	num [1:679]	5.1
DAP	num [1:679]	11.7
Nomecientifico	Factor w/ 65 lev	

O comando 'summary' resume os dados da variável aleatória, se for qualitativa mostra a frequência absoluta das categorias e se for quantitativa apresenta medidas descritivas. Por exemplo:

```
> summary(Nomecientifico)
```

Alsophila sp.	5	Araucaria angustifolia	12	Banara parviflora	1
Blepharocalyx salicifolius	46	Calypttranthes concinna	12	Campomanesia rhombea	7
Campomanesia rhombea	2	Campomanesia rhombea	8	Campomanesia xanthocarpa	19
Casearia decandra	43	Cinnamomum glaziovii	3	Cinnamomum glaziovii	27
Cinnamomum glaziovii	1	Cinnamomum glaziovii	6	Cipós	3
Cryptocarya aschersoniana	27	Cryptocarya aschersoniana	2	Cryptocarya moschata	3
Cupania vernalis	1	Dasyphyllum spinescens	1	Dicksonia sellowiana	10
Eugenia involucrata	9	Eugenia psidiiflora	38	Eugenia psidiiflora	1
Eugenia uruguayensis	48	Eugenia uruguayensis	2	Eugenia uruguayensis	2
Gordonia acutifolia	1	Ilex brevicuspis	39	Ilex brevicuspis	1
Ilex paraguariensis	16	Lamanonia ternata	11	Lamanonia ternata	2
Matayba elaeagnoides	6	Matayba elaeagnoides	1	Myrceugenia cucullata	23
Myrceugenia cucullata	3	Myrceugenia miersiana	4	Myrceugenia miersiana	1
Myrcia oligantha	3	Myrcianthes gigantea	2	Myrciaria delicatula	1
Myrciaria floribunda	20	Myrciaria floribunda	5	Myrciaria tenella	2
Myrsine umbellata	6	Myrsine umbellata	1	Nectandra megapotamica	31
Ocotea indecora	3	Ocotea puberula	1	Ocotea pulchella	21
Podocarpus lambertii	5	Prunus myrtifolia	3	Prunus myrtifolia	1
Rollinia rugulosa	3	Roupala brasiliensis	3	Sapium glandulatum	9
Scutia buxifolia	2	Sebastiania brasiliensis	1	Sebastiania commersoniana	68
Siphoneugena reitzii	5	Syphoneugena reitzii	32	Vernonia discolor	2
Weinmania paulliniifolia	1	Zanthoxylum rhoifolium	1		

```
> summary(DAP)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
9.70 38.05 55.00 73.11 94.95 253.90
```

Observa-se que para 'Nomecientifico' que é uma variável aleatória qualitativa, o resultado apresentado foi as frequências absolutas de cada categoria.

E para 'DAP' que é uma variável aleatória quantitativa, o resultado são as medidas descritivas.

O comando 'tapply' agrega os dados por nível da variável qualitativa. Por exemplo, para saber o DAP médio para cada categoria da variável 'Nomecientifico':

```
> tapply(DAP, Nomecientifico, mean)
```

```
f m
21.30000 21.80822
```

Resultado:

Alsophila sp.	Araucaria angustifolia	Banara parviflora
12.02000	25.32500	21.30000
Blepharocalyx salicifolius	Calyptranthes concinna	Campomanesia rhombea
50.80435	44.27500	58.84286
Campomanesia rhombea	Campomanesia rhombea	Campomanesia xanthocarpa
57.25000	71.17500	67.24211
Casearia decandra	Cinnamomum glaziovii	Cinnamomum glaziovii
40.83488	92.33333	83.34444
Cinnamomum glaziovii	Cinnamomum glaziovii	Cipós
30.20000	33.45000	33.83333
Cryptocarya aschersoniana	Cryptocarya aschersoniana	Cryptocarya moschata
108.54074	118.05000	73.83333
Cupania vernalis	Dasyphyllum spinescens	Dicksonia sellowiana
32.50000	48.10000	66.84000
Eugenia involucreta	Eugenia psidiiflora	Eugenia psidiiflora
53.91111	49.08684	32.80000
Eugenia uruguayensis	Eugenia uruguayensis	Eugenia uruguayensis
55.12917	31.45000	55.05000
Gordonia acutifolia	Ilex brevicuspis	Ilex brevicuspis
53.20000	151.96923	31.00000
Ilex paraguariensis	Lamanonia ternata	Lamanonia ternata
54.33750	120.70909	140.25000
Matayba elaeagnoides	Matayba elaeagnoides	Myrceugenia cucullata
84.56667	35.30000	43.38261
Myrceugenia cucullata	Myrceugenia miersiana	Myrceugenia miersiana
49.80000	41.40000	50.80000
Myrcia oligantha	Myrcianthes gigantea	Myrciaria delicatula
32.26667	87.90000	41.00000
Myrciaria floribunda	Myrciaria floribunda	Myrciaria tenella
48.43000	67.12000	34.50000
Myrsine umbellata	Myrsine umbellata	Nectandra megapotamica
38.33333	33.70000	120.42258
Ocotea indecora	Ocotea puberula	Ocotea pulchella
112.50000	34.70000	130.63810
Podocarpus lambertii	Prunus myrtifolia	Prunus myrtifolia
123.32000	47.33333	35.00000
Rollinia rugulosa	Roupala brasiliensis	Sapium glandulatum
32.10000	67.80000	108.54444
Scutia buxifolia	Sebastiania brasiliensis	Sebastiania commersoniana
128.65000	31.70000	80.31029
Siphoneugena reitzii	Siphoneugena reitzii	Vernonia discolor
75.98000	72.40937	88.45000
Weinmania paulliniifolia	Zanthoxylum rhoifolium	
126.50000	47.40000	

Se um registro possui NA, isto é, dados missing (perdidos):

```
> tapply(DAP, Nomecientifico, mean, na.rm=T)
```

Quando usamos o parâmetro na.rm=T, indicamos para o comando ignorar os NAs nos dados e calcular a média.

O comando subset() é usado para trabalhar com um subconjunto de dados. Por exemplo: só com dados da 'Alsophila sp.':

```
> alsophila <- subset(arvores, Nomecientifico== 'Alsophila sp.')
```

Assim, foi criado um novo objeto que contém só os dados do sexo feminino. É possível combinar vários critérios para criar os subgrupos. Por exemplo, o 'summary' para idade, neste caso:

```
> summary(alsophila$DAP)
```

Por exemplo, para criar um subgrupo com os dados de todas as espécies da pesquisa MENOS da espécie Scutia Buxifolia e com DAP acima de 180 cm:

```
> novoSub <- subset(arvores, Nomecientifico != 'Scutia Buxifolia' & DAP>180)
```

Por exemplo, o 'summary' para DAP, neste novo subconjunto:

```
> summary(novoSub$DAP)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
181.0	187.2	198.3	207.1	218.0	253.9

Operadores booleanos

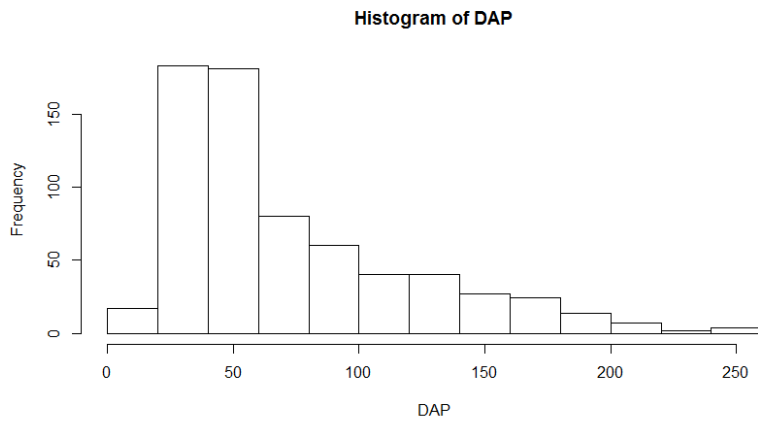
==, !=, >, etc

Gráficos

Vamos elaborar um histograma do DAP:

```
> hist(DAP)
```

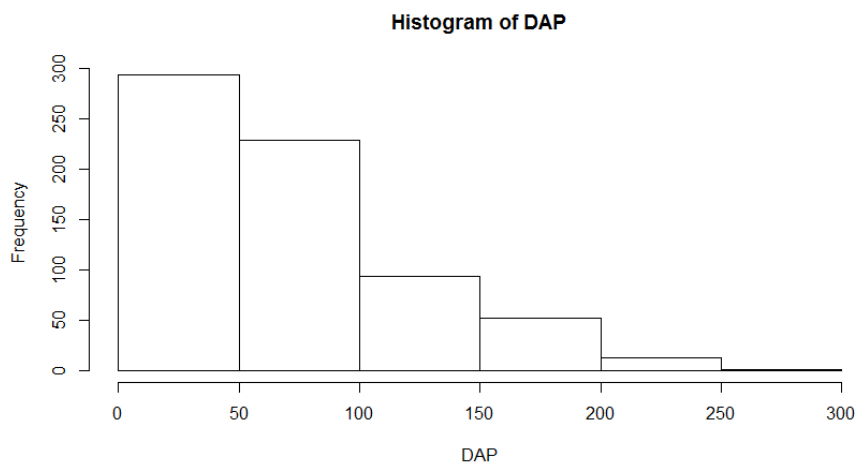
Resultado:



Podemos definir o número de intervalos (colunas) do histograma:

```
> hist(DAP,breaks=5)
```

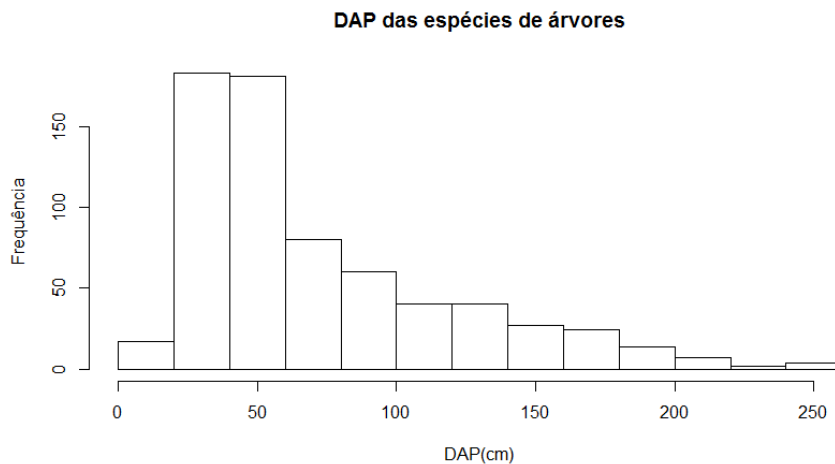
Resultado:



Por padrão, o R coloca os nomes dos eixos em inglês. Para alterar os nomes dos eixos:

```
> hist(DAP, ylab='Frequência', xlab='DAP(cm)', main='DAP das e
spécies de árvores')
```

Resultado:



Entendendo o comando:

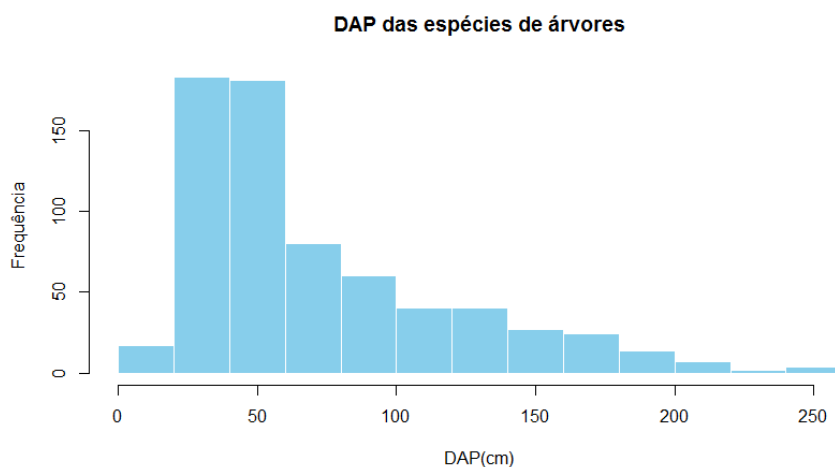
- 'breaks' controla o número de colunas do histograma;
- 'ylab' define o nome do eixo y;
- 'xlab', define o nome do eixo x;
- 'main' define o título do histograma.

Esses parâmetros servem para os outros comandos(tipos) de gráficos.

O parâmetro 'col' controla a cor de preenchimento da coluna e 'border' controla a cor da borda das colunas do histograma, sendo definido por número ou por nome, em inglês. Por exemplo:

```
> hist(DAP, ylab='Frequência', xlab='DAP(cm)', main='DAP das e
    espécies de árvores',col='skyblue',border='white')
```

Resultado:



Podemos ter interesse em sobrepor uma curva sobre o histograma. Por exemplo, para comparar com uma distribuição teórica esperada:

```
> hist(DAP, ylab='Frequência', xlab='DAP(cm)', main='DAP das e  
  espécies de árvores',col='skyblue',border='white',prob=T)
```

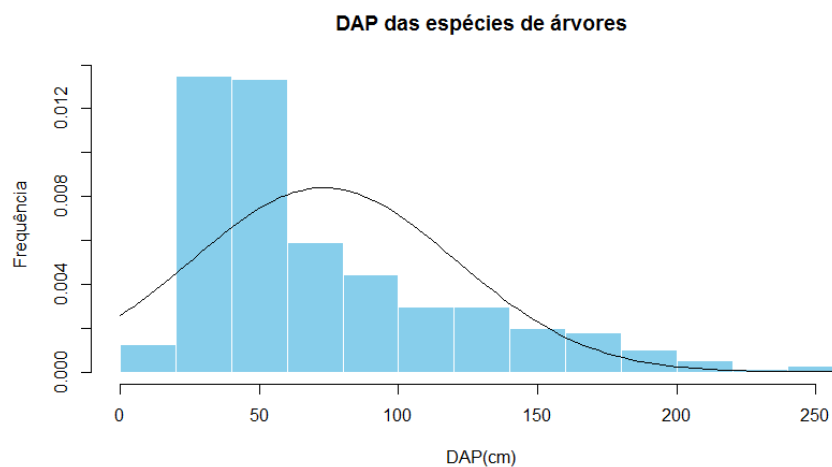
Inserir a curva normal no histograma:

```
> curve(dnorm(x, mean(DAP, na.rm=T), sd(DAP, na.rm=T)), add=T)
```

Entendendo o comando:

- 'dnorm' especifica a densidade de probabilidade da curva teórica que queremos plotar;
- 'x' é uma referência aos valores do eixo x; mean(...) computa a média amostral para usar como parâmetro do dnorm e sd(...) computa o desvio padrão amostral, também para servir de parâmetro para dnorm;
- a opção 'na.rm' nesse comando indica para desconsiderar missing, caso exista;
- 'add=T' força a curva a ser adicionada no gráfico ativo.

Resultado:



2. Estatística Descritiva

2.1 Exemplo

Suponha uma pesquisa (dados simulados) realizada com os alunos da UFFS campus Cerro Largo, em maio de 2016, para saber a opinião quanto a relação ambiente e saúde. Foi utilizada uma amostra aleatória simples de 153 alunos. O cálculo do tamanho da amostra considerou erro de 7%, nível de 94% de confiança, $p=0,5$ e $N=1000$. O instrumento de coleta de dados consta no Apêndice A. O arquivo do banco de dados é disponibilizado na forma digital em formato de planilha eletrônica LibreOffice Calc.

	A	B	C	D	E	F	G	H	I	
1	Numero	idade	sexo	curso	q4	q4sim	q5	q6	q8	q10
2	1	25	m	COMP	s		3 s	n	n	s
3	2	20	f	COMP	s		3 s	s	n	n
4	3	19	f	COMP	s		4 s	n	s	s
5	4	18	m	COMP	s		3 n	n	n	s
6	5	17	f	COMP	s		2 s	n	s	s
7	6	18	m	COMP	s		3 s	n	n	s
8	7	19	m	COMP	s		2 n	n	n	s
9	8	20	f	COMP	s		3 s	n	n	s
10	9	18	m	COMP	s		3 n	n	n	s
11	10	26	m	COMP	s		3 n	n	n	s
12	11	20	f	COMP	s		3 n	n	n	s
13	12	21	m	COMP	s		3 s	n	s	s
14	13	22	f	COMP	s		3 s	n	s	s

Ler os dados no 'Environment' pelo 'Import Dataset...From Local File'.

Para colocar comentários na janela 'console' usamos #.

2.2 Tabelas de frequência

Para elaborarmos uma tabela da variável 'curso' executamos a função `table(curso)`, uma vez que o comando `table` é usado para obter a frequência absoluta de variáveis, ou seja, elaborar as tabelas:

```
> table(curso)
curso
COMP DIR MAT QIA
  27  38  40  48
```

Se desejarmos determinar as frequências relativas:

```
> prop.table(table(curso))
curso
      COMP      DIR      MAT      QIA
0.1764706 0.2483660 0.2614379 0.3137255
```

Para elaborar uma tabela cruzada entre duas variáveis com frequências absolutas, no caso a variável sexo e curso:

```
> table(sexo,curso)
      curso
sexo COMP DIR MAT QIA
f    13  19  24  24
m    14  19  16  24
```

O comando usado para obter uma tabela cruzada com frequências relativas (considerando o total geral):

```
> prop.table(table(sexo,curso))
      curso
sexo COMP      DIR      MAT      QIA
f 0.08496732 0.12418301 0.15686275 0.15686275
m 0.09150327 0.12418301 0.10457516 0.15686275
```

Para obter a tabela de frequências envolvendo a frequência relativa, considerando o total por linha.

```
> prop.table(table(sexo,curso),margin=1)
      curso
sexo COMP      DIR      MAT      QIA
f 0.1625000 0.2375000 0.3000000 0.3000000
m 0.1917808 0.2602740 0.2191781 0.3287671
```

Para obter a tabela de distribuição de frequências envolvendo a frequência relativa, considerando o total por coluna.

```
> prop.table(table(sexo,curso),margin=2)
      curso
sexo COMP      DIR      MAT      QIA
f 0.4814815 0.5000000 0.6000000 0.5000000
m 0.5185185 0.5000000 0.4000000 0.5000000
```

2.3 Medidas descritivas

O comando `tapply(idade,curso,mean)` permite obter a média de idade de acordo com curso. Em outras palavras, usamos este comando para agregar os dados por nível da variável qualitativa.

```
> tapply(idade,curso,mean)
      COMP      DIR      MAT      QIA
20.51852 22.71053 21.50000 21.22917
```


O comando `tapply(idade,curso,sd)` comando que permite obtermos o desvio-padrão por curso.

```
> tapply (idade,curso,sd)
      COMP      DIR      MAT      QIA
2.736792 3.135739 3.522819 2.699406
```

O comando `summary(...)` permite obtermos simultaneamente um conjunto de medidas descrevendo o valor mínimo (Min.), o primeiro quartil (1st Qu.), Mediana (Median), media (Mean) e Quartil 3 (3rd Qu.) e o valor máximo (Max.). Por exemplo para a idade:

```
> summary(idade)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 16.00  20.00   21.00   21.54   23.00   35.00
```

O comando `var(...)` retorna a variância amostral. Por exemplo para a idade:

```
> var(idade)
[1] 9.670881
```

O comando `sd(...)` retorna o desvio-padrão amostral. Por exemplo para a idade:

```
> sd(idade)
[1] 3.109804
```

Para determinar a amplitude total de um conjunto de dados, utilizamos: `max(...)-min(...)`. Por exemplo para a idade:

```
> max(idade)-min(idade)
[1] 19
```

O comando `quantile(...,0.1)` permite determinar o percentil, neste caso o percentil 10. Se utilizarmos 0.2, determinamos o percentil 20.

```
> quantile(idade,0.1)
10%
 18
> quantile(idade,0.2)
20%
 19
```

O comando `subset(table(...),table(...)==max(table(...)))` permite encontrar a moda. O primeiro valor encontrado, refere-se ao valor da moda ao passo que o segundo valor representa quantas vezes esse valor aparece. Por exemplo para a idade:

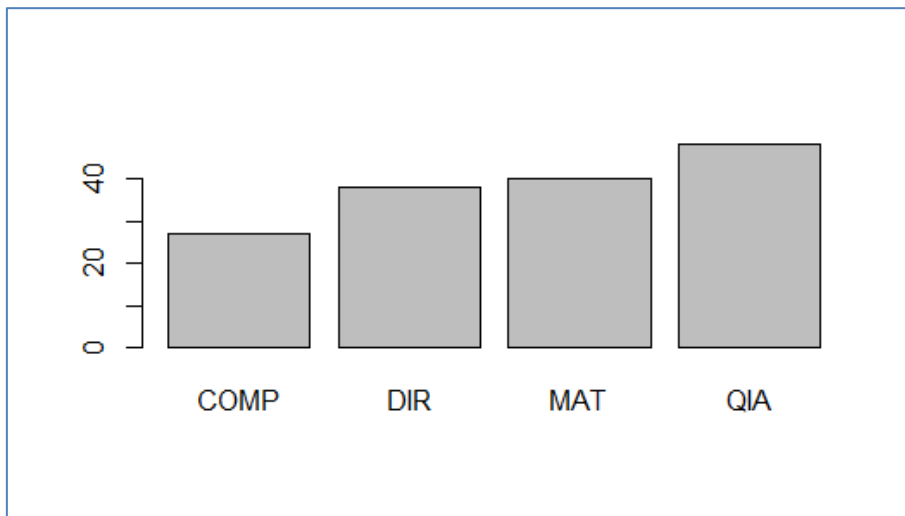
```
> subset(table(idade),table(idade)==max(table(idade)))
20
34
```

2.4 Gráficos

O comando `plot(...)` permite esboçar o gráfico de coluna quando a variável for qualitativa.

```
> barplot(table(curso))
```

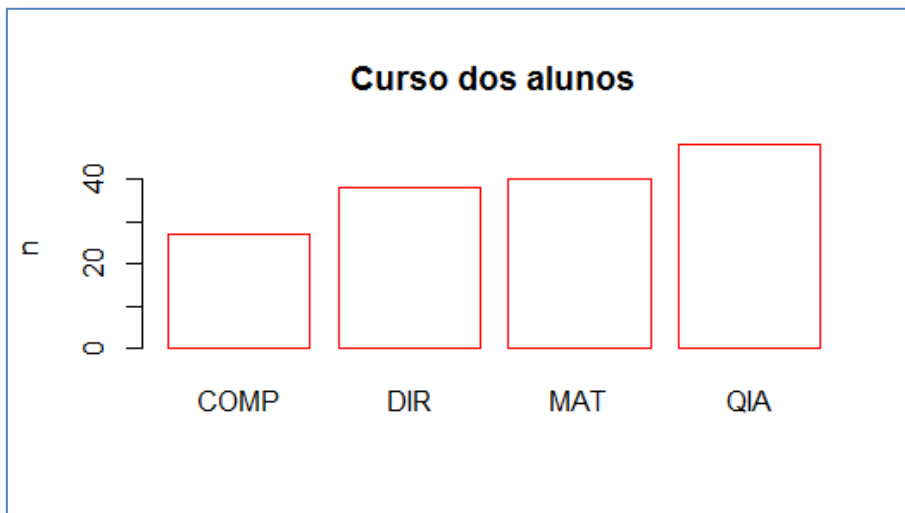
Resultado:



Para personalizar o gráfico:

```
> barplot(table(curso),ylab='n',main='Curso dos alunos',col='white',border='red')
```

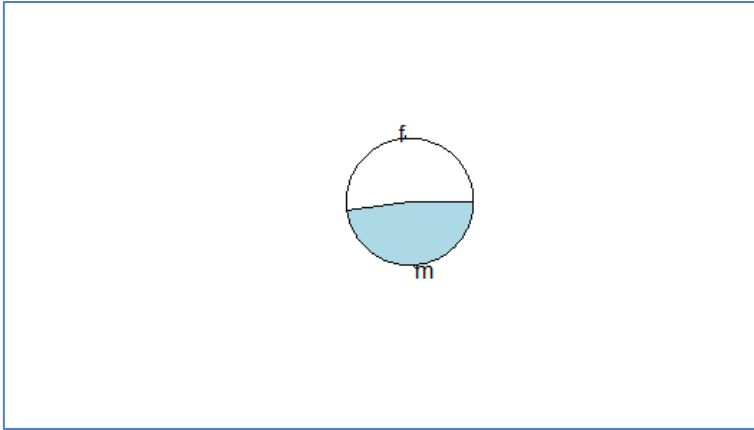
Resultado:



O comando `pie(table(...))` permite elaborar um gráfico de setores. Exemplo para variável sexo:

```
> pie(table(sexo))
```

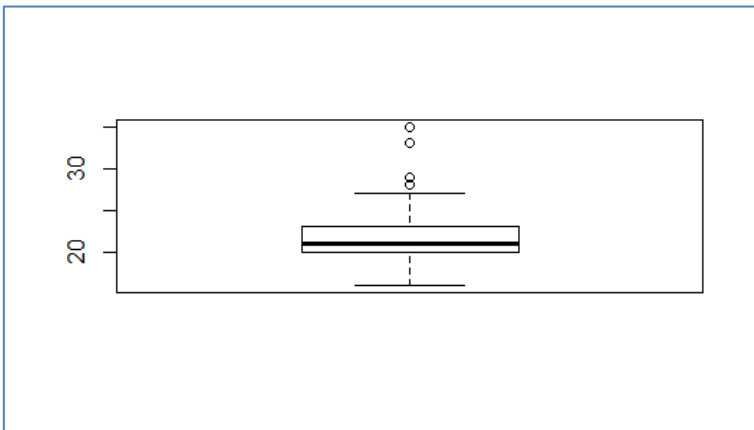
Resultado:



Para obter um boxplot para a variável Idade:

```
> boxplot(idade)
```

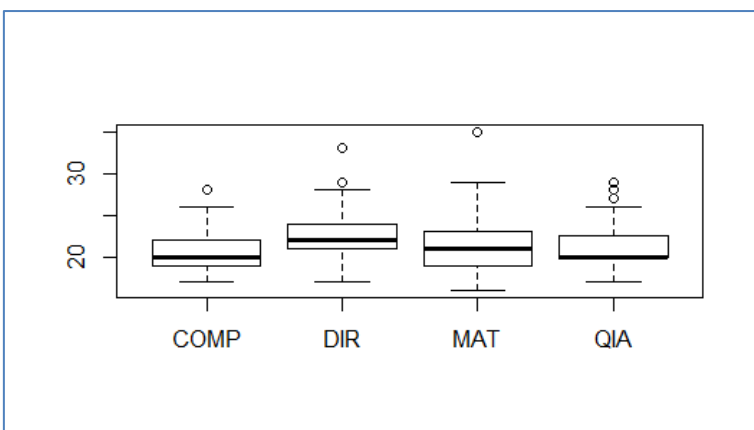
Resultado:



O comando `boxplot(idade~curso)` permite obter o boxplot estratificado por curso.

```
> boxplot(idade~curso)
```

Resultado:

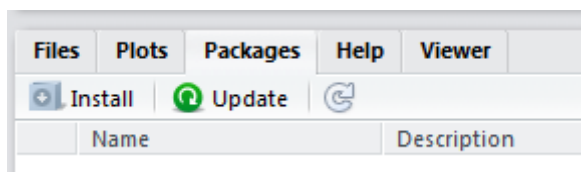


3 Inferência Estatística

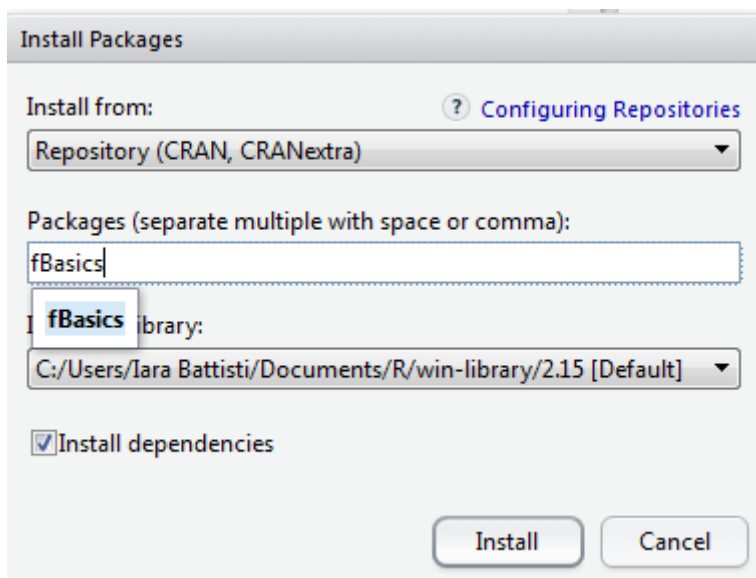
3.1 Intervalo de confiança para média

Para calcular o intervalo de confiança para média de uma população utilizamos o comando 'basicStats' do pacote fBasics. Este pacote é necessário para o cálculo de intervalo de confiança para média.

Assim teremos que instalar o pacote fBasics clicando no botão 'Install' na ficha 'Packages', conforme segue:

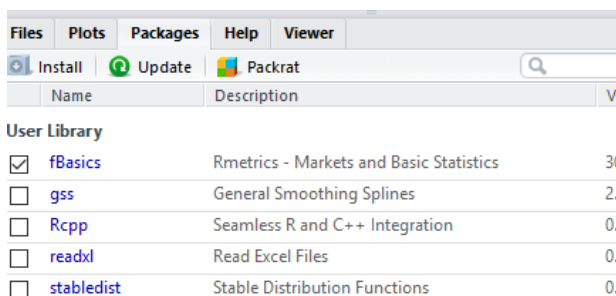


e digitar o nome no pacote 'fBasics' no campo 'Packages', conforme segue:



Após ativar o pacote na linha de comando do painel 'console': `library(fBasics)`

Ou



Comando para calcular o intervalo de confiança para a média de uma população, supondo **normalidade dos dados** e que os dados provem de uma **amostra aleatória simples**:

```
> basicStats(idade,ci=0.95)
      idade
nobs      153.000000
NAS        0.000000
Minimum    16.000000
Maximum    35.000000
1. Quartile 20.000000
3. Quartile 23.000000
Mean       21.542484
Median     21.000000
Sum        3296.000000
SE Mean     0.251413
LCL Mean    21.045769
UCL Mean    22.039198
Variance     9.670881
Stdev       3.109804
Skewness    1.276624
Kurtosis    2.416232
```

em que:

idade: a variável que estamos analisando, isto é, que queremos calcular o intervalo de confiança;

ci: identifica o nível de confiança considerado para o cálculo do intervalo de confiança. O default do R é nível de confiança = 95%.

Os valores estimados para os limites inferior (LCL Mean) e superior do intervalo de confiança (UCL Mean) são 21,05 e 22,04, respectivamente. Indicando que este intervalo contém a média da idade (em anos) dos alunos da população com 95% de confiança (probabilidade).

Este comando também pode ser utilizado para obter estatísticas, semelhante ao comando 'summary'.

3.2 Intervalo de confiança para proporção

Para estimarmos por intervalo proporções binomiais podemos utilizar a aproximação normal em grandes amostras e o intervalo de confiança exato (Ferreira, 2013). Estes métodos são disponibilizados no pacote binom, função 'binom.confint'.

O intervalo de confiança exato para as proporções binomiais deve ser utilizado principalmente se n for pequeno e se p se afastar muito de 1/2 (Ferreira, 2013).

Por exemplo, vamos supor a questão 4 do banco de dados (exemplo) que refere-se a pergunta 'existe saneamento básico em seu município?', neste caso as respostas 'n' refere-se a 'não' e as respostas 's' refere-se a 'sim'.

Assim, o comando para calcular o intervalo de confiança para a proporção da população que neste caso é a proporção de alunos que responderam 'sim' é:

```
> n=153
> y=126
> binom.confint(y,n,conf.level = 0.95, methods = c("exact", "a
symptotic"))
      method x   n      mean      lower      upper
1 asymptotic 126 153 0.8235294 0.7631237 0.8839351
2      exact 126 153 0.8235294 0.7537294 0.8803662
```

Outro exemplo, considerando uma amostra menor:

```
> n <- 48
> y <- 32
> cl <- 0.95
> binom.confint(y,n,conf.level = cl, methods = c("exact",
"asymptotic"))
      method x   n      mean      lower      upper
1 asymptotic 32 48 0.6666667 0.5333080 0.8000253
2      exact 32 48 0.6666667 0.5158917 0.7960403
```

Caso os dados não seguem distribuição normal podemos usar métodos não-paramétricos.

Também, podemos usar para a definição do intervalo de confiança o método de reamostragem, através do comando 'bootstrap'. No bootstrap assumimos que a distribuição empírica da amostra é uma boa aproximação da distribuição populacional da variável de interesse.

3.3 Teste de hipótese

Primeiramente temos que estudar a teoria dos testes de hipóteses e entender a interpretação do valor p.

Erro Tipo I e Erro Tipo II em testes de hipóteses:

Resultados possíveis em um teste de hipótese e suas probabilidades de ocorrência		
Decisão tomada	A verdade na população	
	H_0 é verdadeira	H_0 é falsa
H_0 não é rejeitada	Decisão correta ($1 - \alpha$)	Decisão errada Erro Tipo II - (β)
H_0 é rejeitada	Decisão errada Erro Tipo I - (α)	Decisão correta ($1 - \beta$)

Definição das hipóteses:

Hipótese Nula (H_0): é uma afirmação sobre o valor de um parâmetro populacional, deve conter a condição de igualdade. Exemplo para o teste de média populacional:

$$H_0 : \mu = \text{algum valor}$$

$$H_0 : \mu \geq \text{algum valor}$$

$$H_0 : \mu \leq \text{algum valor}$$

Hipótese Alternativa (H_1): é a afirmação que deve ser verdadeira se a hipótese nula é falsa. Para a média, a hipótese alternativa comporta uma das três formas:

$$H_1 : \mu \neq \text{algum valor}$$

$$H_1 : \mu < \text{algum valor}$$

$$H_1 : \mu > \text{algum valor}$$

Teste Bilateral, Unilateral Direito, Unilateral Esquerdo: rejeita-se H_0 , se o valor calculado está na região de rejeição (área pintada dos gráficos da Figura 1), porque isto indica uma discrepância significativa entre H_0 e os dados amostrais.

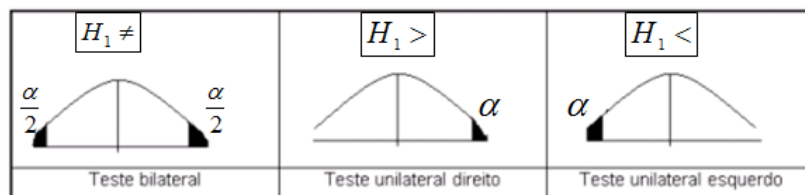


Figura 1 – Teste bilateral, unilateral direito e unilateral esquerdo

Valor p:

O **valor p** reflete a plausibilidade de se obter tais resultados no caso de H_0 ser de fato verdadeira

Regra de decisão do valor p

$p \leq 0,01$. Rejeita-se H_0 ao nível de 1% de significância.

$0,01 < p \leq 0,05$. Rejeita-se H_0 ao nível de 5% de significância.

$P > 0,05$. Não rejeita-se H_0 .

3.3.1 Teste de hipótese para verificar a normalidade dos dados

Para verificar se os dados seguem uma distribuição normal, podemos inicialmente usar o histograma e depois confirmar com um teste estatístico para testar normalidade como Shapiro-Wilk.

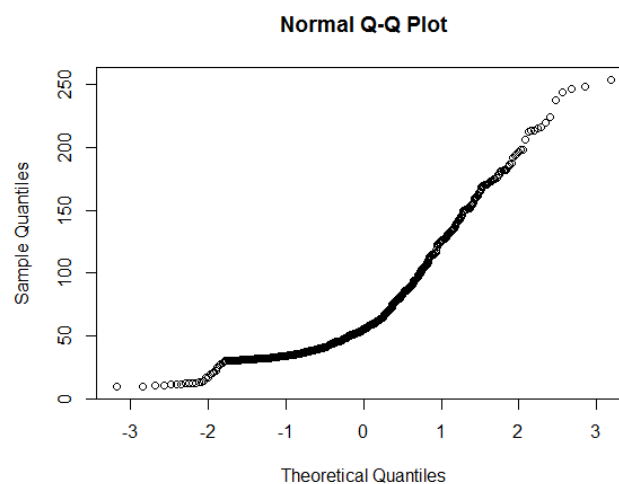
```
> shapiro.test(idade)

Shapiro-wilk normality test

data:  idade
W = 0.90677, p-value = 2.536e-08
```

Obtendo o gráfico de probabilidade normal:

```
➤ qqnorm(idade)
```



3.4 Teste de hipótese para verificar homogeneidade de variâncias

Teste F para teste homogeneidade de variância:

```
> var.test(idade~q4)
```

F test to compare two variances

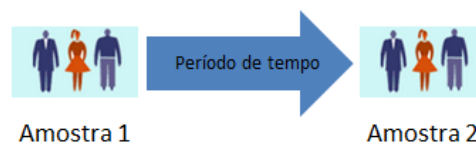
```
data:  idade by q4
F = 0.64492, num df = 26, denom df = 125, p-value = 0.1944
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3730846 1.2585858
sample estimates:
ratio of variances
 0.6449208
```

Outro teste para verificar a homogeneidade de variâncias, neste caso é necessário instalar o pacote 'car' e ativa-lo.


```
> leveneTest(idade~q4,center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  1.7287 0.1906
151
```

3.5 Teste de hipótese para comparação de duas médias entre duas amostras dependentes

Também chamadas de amostras emparelhadas ou pareadas, se uma amostra tem relação com a outra.



Para exemplificar, vamos digitar os dados diretamente na linha de comando:

Exemplo: É obtido o peso de seis indivíduos antes e após um treinamento de exercício físico. Teste a hipótese de que a média antes do treinamento é diferente da média após o treinamento. Utiliza o nível de significância de 0,05.

Indivíduo	A	B	C	D	E	F
Peso antes do treinamento (kg)	99	62	74	59	70	73
Peso após o treinamento (kg)	94	62	66	58	70	76

```
> antes=c(99,62,74,59,70,73)
> depois=c(94,62,66,58,70,76)
```

```
t.test(antes,depois,paired=TRUE)
```

Paired t-test

```
data: antes and depois
t = 1.131, df = 5, p-value = 0.3094
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.333688  6.000355
sample estimates:
mean of the differences
      1.833333
```

1º passo) Definir hipóteses

H0: média antes = média depois
H1: média antes ≠ média depois

2º passo) Aplicar o teste estatístico adequado

Como são amostras dependentes, utiliza-se o teste t para comparar amostras dependentes, assim:

- o primeiro parâmetro dentro do parênteses define a amostra antes;
- o segundo parâmetro dentro do parênteses define a amostra depois;
- paired=TRUE define que são duas amostras dependentes.

3º passo) Conclusão (valor p = p value)

p-value = 0.3094

Não rejeita-se H_0 . Portanto, a média do peso corporal antes do treinamento de exercício físico é igual a média do peso corporal depois do treinamento de exercício físico.

Ou

Os dados amostrais não forneceram evidências suficientes para rejeitar a hipótese nula. Portanto, as médias dos pesos corporais antes e depois do treinamento são iguais.

3.6 Teste de hipótese para comparação de duas médias entre duas amostras independentes

Duas amostras são independentes se a amostra extraída de uma das populações não tem qualquer relação com a amostra extraída da outra população.



Amostra 1



Amostra 2

Primeiramente precisamos saber se existe homogeneidade de variâncias populacionais, a qual poderá ser verificada através de um teste de homogeneidade de variâncias utilizando os dados das duas amostras.

Como exemplo, suponha que queremos comparar a idade média entre os alunos que responderam 'sim' e os alunos que responderam 'não' na questão 4.

1º passo) Definir hipóteses

H_0 : média de idade q4sim = média de idade q4não

H_1 : média de idade q4sim \neq média de idade q4não

2º passo) Aplicar o teste estatístico adequado

Teste t para duas amostras independentes (paired=FALSE) e variâncias populacionais iguais (var.equal=TRUE).

```
> t.test(idade~q4,var.equal=TRUE,paired=FALSE)
Two Sample t-test
```

```
data: idade by q4
t = -0.4521, df = 151, p-value = 0.6518
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.605395  1.007512
sample estimates:
mean in group n mean in group s
    21.29630      21.59524
```

3º passo) Conclusão (valor $p = p\text{ value}$)

$p > 0,05$

Não rejeita-se H_0 . Portanto, não há evidências para afirmar que a média de idade é diferente entre os alunos que responderam 'sim' e os alunos que responderam 'não' na questão 4.

REVISANDO:

Quando as amostras são **independentes**, devemos primeiramente testar se as variâncias populacionais são iguais ou diferentes, usando o teste `var.test` ou `levene.test`.

Se o resultado do teste ($p < 0,01$) for altamente significativo ou ($0,01 < p < 0,05$) significativo então concluímos que as variâncias são diferentes, desta forma devemos indicar no comando `var.equal=FALSE`.

No caso do teste não se significativo ($p > 0,05$) então concluímos que as variâncias não são diferentes, desta forma devemos indicar no comando `var.equal=TRUE`.

4. Referências

- TRIOLA, M. F. **Introdução à estatística**. 2. ed. Rio de Janeiro: LTC, 1999.
FERREIRA, D. F. **Recursos Computacionais Utilizando o R**. UFLA, 2013.

APÊNDICE A

INSTRUMENTO DE COLETA DE DADOS

Indicadores de Saúde e Meio Ambiente e sua inter-relação na Região das Missões

O objetivo da pesquisa é conhecer a percepção dos alunos da UFFS, campus Cerro Largo, em maior de 2015, através de uma amostra aleatória. Agradecemos a sua participação!

DADOS DE CARACTERIZAÇÃO

Município Residente: _____

Idade: ____

Sexo: () Masculino () Feminino

Renda familiar (SM): _____

Grau de escolaridade:

- () Não alfabetizado () Ensino fundamental incompleto () Ensino fundamental completo
() Ensino médio incompleto () Ensino médio completo () Ensino superior incompleto
() Ensino superior completo () Pós-graduação

SANEAMENTO E SAÚDE

Questão 1. Para você, qual é a relação existente entre meio ambiente e saúde?

Questão 2. Para você, o que é saneamento básico?

Questão 3. Em sua opinião, como o saneamento básico pode interferir na saúde da população?

Questão 4. Existe saneamento básico em seu município?

() Sim () Não

Se sim, como você avalia o saneamento básico no seu município?

() Ótimo () Bom () Regular () Ruim () Péssimo

Se não, como você avalia a falta deste serviço?

Questão 5. Você acha que no seu município existem casos de doenças relacionadas ao saneamento básico?

() Sim () Não

Questão 6. Na sua residência, já houve casos de doenças causadas pela falta de saneamento?

() Sim () Não

Se sim, quais doenças?

Se sim, já houve óbitos de crianças menores de 5 anos de idade devido a alguma doença relacionada ao saneamento em sua residência?

() Sim () Não

RESÍDUO SÓLIDO

Questão 7. O que é resíduo (lixo) para você?

Questão 8. Você costuma separar os resíduos em sua casa?

☐ Sim ☐ Não

Se sim, como? _____

Questão 9. Para você, como a coleta de resíduos (sólidos e orgânicos) tem influência na saúde da população?

Questão 10. Em sua residência, existe a coleta de resíduos?

☐ Sim ☐ Não

Se sim, qual é a frequência de coleta de resíduos no seu domicílio?

Se sim, como você avalia a coleta de resíduos em seu município?

☐ Ótima ☐ Boa ☐ Regular ☐ Ruim ☐ Péssima

Questão 11. Você sabe qual o local em que são depositados os resíduos após a coleta em seu município?

☐ Sim ☐ Não.

Se sim, onde? _____

Questão 12. Em sua opinião, qual seria o local adequado para a disposição dos resíduos gerados pela população do município?

Questão 13. Para você, existe diferença entre aterro sanitário e lixão?

☐ Sim ☐ Não

Se sim, qual? _____

Questão 14. Existe a coleta seletiva de resíduos em seu município?

☐ Sim ☐ Não

Questão 15. Para você, qual é a importância da coleta seletiva?

Questão 16. Você acha que teria maneiras de aproveitar os resíduos sólidos e orgânicos?

☐ Sim ☐ Não

Se sim, quais maneiras seriam?

ÁGUA

Questão 17. Você acha que a qualidade da água influencia na vida da população?

☐ Sim ☐ Não

Se sim, como? _____

Questão 18. Em sua residência, a água é provida da rede de abastecimento pública?

☐ Sim ☐ Não

Se não, de onde provém a água de sua residência?

Questão 19. A sua residência possui caixa d'água?

☐ Sim fechada com tampa ☐ Sim mantida aberta ☐ Não

Se sim, é feita a limpeza desta? ☐ Sim ☐ Não

Se sim, com que frequência? _____

Questão 20. Ocorre falta de água em sua residência?

☐ Sim ☐ Não

Se sim, avalie no intervalo de 1 a 5, a frequência dessa falta de água. (1 – não há falta e 5 – falta muito frequente)

() 1 () 2 () 3 () 4 () 5

Questão 21. No seu município existe algum tratamento de água para consumo da população?

() Sim () Não () Não sei

Se sim, qual? _____

Questão 22. Em sua residência, já houve casos de doenças relacionadas com a má qualidade da água?

() Sim () Não

Se sim, quais? _____

Questão 23. Como você avalia a qualidade da água em seu município?

() Ótima () Boa () Regular () Ruim () Péssima

ESGOTO

Questão 24. Para você, como a coleta de esgoto tem influência na qualidade de vida da população?

Questão 25. O seu município possui rede pública de esgoto?

() Sim () Não

Questão 26. A sua residência está ligada à rede pública de esgoto?

() Sim () Não

Se não, onde é depositado o esgoto proveniente de sua residência?

() poço negro () fossa séptica () outro _____

Questão 27. Para você, existe diferença entre poço negro e fossa séptica?

() Sim () Não

Se sim, qual? _____

Questão 28. Em sua opinião, qual é a melhor forma de descarte para o esgoto gerado em sua residência?

Questão 29. Em sua residência, já houve casos de doenças relacionadas ao descarte de esgoto inadequado?

() Sim () Não

Se sim, quais? _____

Questão 30. Como você avalia a coleta de esgoto em seu município?

() Ótima () Boa () Regular () Ruim () Péssima

VETORES

Questão 31. A sua residência possui incidência de:

Mosca () sim () não Pernilongo () sim () não

Borrachudo () sim () não Barata () sim () não

Rato () sim () não Formiga () sim () não

Outros _____

Questão 32. Para você, qual é a possível relação existente entre esses vetores (questão 31) com a saúde da população? _____

Questão 33. Você toma algumas ações para prevenir a entrada desses vetores (questão 31) em sua residência?

() Sim () Não

Se sim, quais? _____

Questão 34. Em sua residência, já houve casos de doenças relacionadas a algum desses vetores (questão 31)?

() Sim () Não

Se sim, quais? _____

QUESTÕES GERAIS

Questão 35. Os habitantes de sua residência possuem algum plano de saúde?

() Sim () Não

Se sim, este é efetivo quando há necessidade? () Sim () Não

Questão 36. No geral, como você avalia a qualidade de vida da população do seu município e de sua própria residência?

Referente ao Município: () Ótima () Boa () Regular () Ruim () Péssima

Referente à Residência: () Ótima () Boa () Regular () Ruim () Péssima

Questão 37. Ocorre regularmente a visita de agentes sanitários em sua residência?

() Sim () Não

Sem sim, quando foi à última visita?

Como ocorreram e por quais motivos?

Questão 38. O que você acha que poderia ser feito pela população e pelos órgãos públicos para melhorar a relação entre ambiente e saúde em seu município?

Questão 39. Existem alguns planos/projetos sobre a relação saúde e ambiente que você conhece?
