Fundamentos de Estatística para Ciência de Dados

Filipe J. Zabala

2020-11-07

Contents

Prefácio

Há dois motivos para ler este texto: (*i*) você deseja se tornar um profissional qualificado e entende que o ferramental estatístico pode auxiliar em suas decisões futuras ou (*ii*) você foi obrigado. De toda forma sugere-se a leitura deste[1][2] e de outros materiais de apoio ao longo do curso praticando através de exercícios de fixação. Este texto consiste em declarações tão verídicas quanto o possível para a linguagem humana usual.

Jingle do Livro[3]

Leia este livro
É uma pesquisa paciente
Cada linha deste texto
No papel ou virtualmente
Fará você ficar
Ao menos inteligente

[1] Este material foi desenvolvido nos ambientes RStudio v1.4.904 e R 4.0.3 baseado no pacote bookdown, disponível em `www.filipezabala.com`.

[2] Este material está sob a licença Creative Commons de Atribuição 4.0 Internacional (CC BY 4.0). Você tem o direito de compartilhar – copiar e redistribuir o material em qualquer suporte ou formato – e adaptar – remixar, transformar, e criar a partir do material para qualquer fim, mesmo que comercial. Você deve dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas. Você deve fazê-lo em qualquer circunstância razoável, mas de nenhuma maneira que sugira que o licenciante apoia você ou o seu uso. Você não pode aplicar termos jurídicos ou medidas de caráter tecnológico que restrinjam legalmente outros de fazerem algo que a licença permita. Mais detalhes em creativecommons.org/licenses/by/4.0/legalcode.

[3] 'Compre este disco / É uma pesquisa paciente / Cada volta da agulha / Pelo sulco docemente / Fará você ficar / Mais feliz e inteligente'. *Jingle do Disco* de Tom Zé, do álbum *1992 The Hips of Tradition*.

Chapter 1

Introdução

O *Método Estatístico* ou simplesmente *Estatística* reúne ferramentas teóricas e práticas para analisar informações quantitativas, medir incertezas e auxiliar na tomada de decisão. É um componente do *Método Científico*, e pode ser dividido conforme o esquema da Figura a seguir. Neste curso serão abordados tópicos de Estatística Descritiva,, Inferência Estatística sob os prismas da Estatística Clássica (ou Frequentista) e Bayesiana e Séreis Temporais.

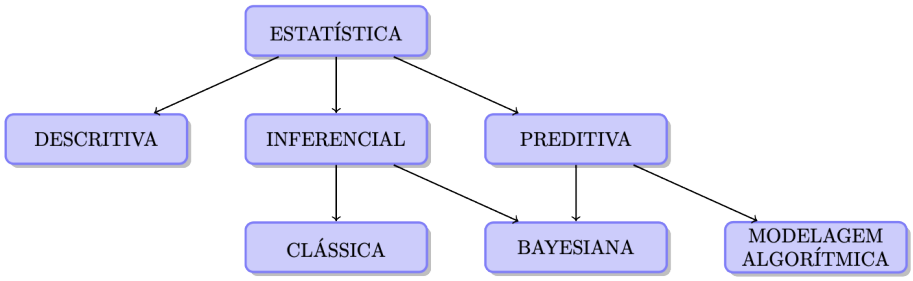


Figure 1.1: Uma possível divisão da Estatística.

1.1 Ferramentas

1.1.1 R

R é uma ferramenta para cálculos estatísticos e gráficos. Foi desenvolvido no departamento de Estatística da Universidade de Auckland, e seu código está disponível sob a licença GNU (*GNU is Not Unix*) GPL[1]. Atualmente a *R Foun-*

[1]A Licença Pública Geral GNU é um tipo de licença utilizada para software livre, que garante aos usuários finais (indivíduos, organizações ou empresas) a liberdade de usar, estudar, compartilhar e modificar o software.

dation está sediada na Universidade de Economia e Negócios de Viena, Áustria. Foi influenciado por linguagens como *S* e *Scheme* seguindo o conceito minimalista orientado a objeto, que especifica um pequeno núcleo padrão acompanhado de pacotes para a extensão da linguagem.

Recomenda-se manter o R e seus pacotes sempre atualizados. No Windows recomenda-se ainda a instalação do Rtools de acordo com a versão instalada do R. Os pacotes utilizados neste curso podem ser instalados e atualizados conforme código abaixo. No caso de utilização de sistema operacional do tipo Unix, recomenda-se rodar as instruções acima em um terminal após executar o comando `sudo R` seguido da senha do sistema.

```
packs <- c('tidyverse','readxl','e1071','arrangements','DescTools','symmetry',
           'mvtnorm','VGAM','chisq.posthoc.test','rgl','ggfortify','factoextra',
           'reticulate')
install.packages(packs, dep = T)
devtools::install_github('filipezabala/jurimetrics', force=T)
devtools::install_github('filipezabala/voice', force=T)
devtools::install_github('filipezabala/desempateTecnico', force=T)
update.packages(ask = F)
```

CRAN Task Views

As CRAN Task Views visam fornecer informações sobre os pacotes da CRAN (*Comprehensive R Archive Network*) relacionados a um determinado tópico. É recomendado verificar os assuntos de interesse dentro das CRAN Task Views para uma abordagem mais completa utilizando a linguagem R.

1.1.2 RStudio

RStudio é um ambiente de desenvolvimento integrado ao R. Possibilita a criação de apresentações e relatórios automáticos em diversos formatos como pdf, html e docx, mesclando linguagens como R, LaTeX, markdown, C, C++, Python, SQL, HTML, CSS, JavaScript, Stan e D3. Está disponível nas edições Desktop, Server juntamente com seus respectivos *previews*, reunindo as funcionalidades do R de forma parcimoniosa.

1.1.3 Python

Python é uma linguagem de programação interpretada, interativa e orientada a objetos. Ela incorpora módulos, exceções, tipagem dinâmica, tipos de dados dinâmicos de nível muito alto e classes. Oferece suporte a vários paradigmas de programação além da programação orientada a objetos, como a programação procedural e funcional. Ele tem interfaces para muitas chamadas de sistema e bibliotecas, bem como para vários sistemas de janela, e é extensível em C ou C ++. Também pode ser usado como uma linguagem de extensão para aplicativos

que precisam de uma interface programável. Finalmente, o Python é portátil: ele roda em muitas variantes do Unix, incluindo Linux e macOS, e no Windows.

Python em R Markdown

O pacote `reticulate` inclui um mecanismo Python para R Markdown que executa trechos de Python em uma única sessão Python incorporada em sua sessão R, permitindo o acesso a objetos criados em trechos de Python do R e vice-versa.

```
# Interface de R e Python
library(reticulate)
repl_python()
```

```
## Python 3.8.5 (/usr/local/bin/python3.8)
## Reticulate 1.18 REPL -- A Python interpreter in R.
```

```
os <- import("os")
os$listdir(".")
```

```
##  [1] "02-descritiva.Rmd"       ".Rhistory"               "ei_files"                "01-introdu
##  [5] ".DS_Store"               "missfont.log"            "LICENSE"                 "temp.zip"
##  [9] "preamble.tex"            "index.Rmd"               "03-prob.Rmd"             "Dockerfile
## [13] "_deploy.sh"              "latex"                   "07-modelos_lineares.Rmd" "TinyTeX.tg
## [17] "packages.bib"            "code"                    "05-inferencia-class.Rmd" "ei.Rmd"
## [21] "temp.xlsx"               "_output.yml"             "0ei.Rproj"               "_bookdown_
## [25] "06-inferencia-bayes.Rmd" "img"                     "_bookdown.yml"           "DESCRIPTIO
## [29] ".gitignore"              "09-series_temporais.Rmd" "style.css"               "_book"
## [33] "ei.log"                  "source_rfcv.r"           "now.json"                "book.bib"
## [37] "08-apr_maquina.Rmd"      ".git"                    "data"                    ".Rproj.use
## [41] "toc.css"                 "_build.sh"               "10-ref.Rmd"              "04-amostra
```

Exercício 1.1. Ler a documentação do `reticulate` disponível em https://rstudio.github.io/reticulate/.

1.1.4 JASP

JASP é um projeto de código aberto apoiado pela Universidade de Amsterdã. Com interface amigável, oferece procedimentos de análises estatísticas com abordagens clássica e bayesiana. Desenvolvido para análises de publicação, dentre suas principais características, estão:

- Atualização dinâmica de todos os resultados
- Layout de planilha e uma interface intuitiva de arrastar e soltar
- Saída anotada para comunicar seus resultados

- Integração com o *Open Science Framework* (OSF)
- Suporte para formato APA (copie gráficos e tabelas diretamente no Word)

1.1.5 Stan

Stan é uma plataforma de código aberto para modelagem estatística e computação estatística de alto desempenho. É também utilizado para análise de dados e previsão nas ciências sociais, biológicas e físicas, engenharia e negócios. A biblioteca de matemática de Stan fornece funções de probabilidade e álgebra linear. Pacotes de R adicionais fornecem modelagem linear baseada em expressão, visualização da posteriori e validação cruzada de exclusão. Existem interfaces para diversos ambientes de computação populares, tais como RStan (R) e PyStan (Python). Usando a linguagem pode-se obter:

- inferência estatística bayesiana completa com amostragem MCMC (NUTS, HMC)
- inferência bayesiana aproximada com inferência variacional (ADVI)
- estimativa de máxima verossimilhança penalizada com otimização (L-BFGS)

1.2 Materiais de apoio

1.2.1 Página do professor Filipe Zabala

Em filipezabala.com o aluno irá encontrar uma série de materiais de apoio como apostilas, vídeos e artigos. Em github.com/filipezabala estão disponíveis uma série de repositórios criados pelo professor.

Exercício 1.2. Para uma introdução aos conceitos básicos de R, assita aos vídeos disponíveis na playlist Ciência de Dados em software livre.

1.2.2 Khan Academy

A Khan Academy[2] possui uma ampla gama de materiais gratuitos em Português, que podem servir de suporte ao aluno durante o curso. A lista a seguir indica os principais fundamentos necessários para o bom desenvolvimento do conteúdo.

1. Propriedades fundamentais de potenciação, radiciação e frações
- https://pt.khanacademy.org/math/brazil-math-grades/pt-5-ano/numeros-fracoes-5ano

[2]Segundo a informação oficial, 'é uma organização sem fins lucrativos com a missão de oferecer uma educação gratuita de alta qualidade para qualquer pessoa, em qualquer lugar'.

- https://pt.khanacademy.org/math/brazil-math-grades/pt-8-ano/numeros-8ano

2. Teoria dos Conjuntos

- https://pt.khanacademy.org/math/6-ano-matematica/numeros-operacoes-com-numeros-naturais-6ano
- https://pt.khanacademy.org/math/brazil-math-grades/pt-9-ano/numeros-9ano
- https://pt.khanacademy.org/math/statistics-probability/probability-library#basic-set-ops

3. Análise combinatória e axiomas de probabilidade

- https://pt.khanacademy.org/math/brazil-math-grades/pt-7-ano/probabilidade-e-estatistica-7ano
- https://pt.khanacademy.org/math/brazil-math-grades/pt-8-ano/probabilidade-e-estatistica-8ano
- https://pt.khanacademy.org/math/brazil-math-grades/pt-9-ano/probabilidade-e-estistica-9ano

4. Funções elementares: linear, polinomial, logarítmica e exponencial

- https://pt.khanacademy.org/math/brazil-math-grades/pt-9-ano/algebra-funcoes-9ano
- https://pt.khanacademy.org/math/algebra2/exponential-and-logarithmic-functions

5. Matrizes, determinantes, decomposições, autovalores e autovetores

- https://pt.khanacademy.org/math/algebra-home/alg-matrices

6. Derivadas e integrais

- https://pt.khanacademy.org/math/differential-calculus/dc-diff-intro
- https://pt.khanacademy.org/math/calculus-home/integration-calc

1.3 Algarismos e Números

Um *algarismo* é um símbolo, enquanto um *número* expressa uma idéia de quantidade. Números são representados por algarismos, sendo fundamental distinguir estes elementos.

Exemplo 1.1. Se há 20 alunos na sala A outros 30 na sala B, pode-se dizer que, em média, há $\frac{20+30}{2} = 25$ alunos nas duas salas. Esta é uma informação numérica. Se rotularmos o sexo masculino como 0 e o feminino como 1, fica claro que 0 e 1 estão sendo tratados como algarismos, uma vez que não expressam quantidades.

1.4 Porcentagens, Decimais e Milhares

Neste texto será adotado o padrão americano, que utiliza o símbolo de ponto (.) como separador de decimais e vírgula (,) como separador de milhares. Assim,

$$\frac{1}{40} = 0.025 = 0.0250 = .025 = 2.5\% = \frac{2.5}{100}.$$

Dízimas periódicas serão escritas na forma $\frac{1}{3} = 0.333... = 0.\bar{3} \approx 0.333 \approx 0.3$. O número $32,960 = 30,000 + 2,000 + 960$ deve ser lido como 'trinta e dois mil novecentos e sessenta'.
Esta opção evita muitos problemas, já que muitos *softwares* estatísticos não são compatíveis com o padrão brasileiro, que utiliza vírgula como separador de decimais e ponto para separar os milhares. Nas anotações pessoais e listas de exercícios poderá ser adotada a notação de preferência do aluno.

1.5 O Senhor X

Quando avalia-se algo de interesse prático, em geral observam-se nomes longos. Considere a variável

> X: 'número de filhos de mulheres atendidas em um hospital público de Porto Alegre em 2019'.

Esta longa descrição tornará maçante qualquer texto que utilize-o muitas vezes, tornando impraticável a realização de cálculos envolvendo tal característica de interesse. É razoável, portanto, associar descrições longas a símbolos. A letra X é famosa por simbolizar algo genérico, tanto na Ciência quanto na vida cotidiana. Note que o símbolo utilizado para separar X de sua descrição é `:`, e não `=`.

Neste texto será utilizado X (maiúsculo) para representar a característica de interesse, e x_k (minúsculo) para representar o k-ésimo valor observado desta característica. Assim, enquanto X representa genericamente o número de filhos de mulheres atendidas em um hospital público de Porto Alegre em 2019, $x_4 = 2$ indica que a quarta mulher avaliada no estudo tem dois filhos.

1.6 Somatório

A soma de n números $x_1, x_2, ..., x_n$ é representada por $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$, e lê-se 'somatório de xis i de um até ene'.

Exemplo 1.2. (Número de passos) Suponha que foi anotado o 'número de passos até a lixeira mais próxima' na cidade de Porto Alegre em $n = 6$ ocasiões, conforme Tabela a seguir.

x_1	x_2	x_3	x_4	x_5	x_6
186	402	191	20	7	124

Esta tabela indica que na primeira ocasião foram caminhados 186 passos até localizar uma lixeira (representado por $x_1 = 186$), na segunda foram 402 passos (representado por $x_2 = 402$), e assim sucessivamente. Para calcular o total de passos caminhados, pode-se fazer

$$\sum_{i=1}^{6} x_i = x_1 + x_2 + \cdots + x_6 = 186 + 402 + 191 + 20 + 7 + 124 = 930 \quad (1.1)$$

```
186+402+191+20+7+124              # R e RStudio são calculadoras
```

```
## [1] 930
```

```
x <- c(186,402,191,20,7,124)      # Pode-se criar um vetor e atribuir a x
sum(x)                            # Usando a função 'sum', apresentada na Equação (1.1)
```

```
## [1] 930
```

```
sum(x^2)                          # Soma dos quadrados, representada pela Equação (1.2)
```

```
## [1] 248506
```

A letra grega $\sum$ é o sigma maiúsculo, conforme Seção 1.8.1. Em muitos casos a simbologia de somatório é simplificada, utilizando-se $\sum$, $\sum_x$ ou $\sum_i$. A seguir estão alguns exemplos mais avançados de uso mais sofisticado do somatório, podendo ser omitidos em uma primeira leitura.

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + ... + x_n^2 \quad (1.2)$$

Exercício 1.3. Considere o banco de dados disponível no pacote `coronavirus`[3] conforme código a seguir.

```
library(coronavirus)     # chamando a biblioteca 'coronavirus'
# update_dataset(silence = FALSE)  # atualizando os dados
data(coronavirus)        # deixando o banco de dados disponível
dim(coronavirus)         # dimensões do banco de dados (linhas x colunas)
```

[3] *Johns Hopkins University Center for Systems Science and Engineering* (JHU CCSE). https://systems.jhu.edu/research/public-health/ncov

```
## [1] 219852      7
```

```
head(coronavirus)       # mostrando o início do banco de dados
```

```
##         date province     country  lat long      type cases
## 1 2020-01-22          Afghanistan 33.9 67.7 confirmed     0
## 2 2020-01-23          Afghanistan 33.9 67.7 confirmed     0
## 3 2020-01-24          Afghanistan 33.9 67.7 confirmed     0
## 4 2020-01-25          Afghanistan 33.9 67.7 confirmed     0
## 5 2020-01-26          Afghanistan 33.9 67.7 confirmed     0
## 6 2020-01-27          Afghanistan 33.9 67.7 confirmed     0
```

a. Obtenha a soma de casos (`cases`) registrados ao longo de todo o período.
b. Obtenha a soma ao quadrado de casos registrados ao longo de todo o período.
c. Obtenha a soma de casos registrados ao longo de todo o período dividido por tipo (`type`).
d. Considerando a variável X: 'número de casos registrados' em `nrow(coronavirus)` linhas do banco de dados, represente os itens a. e b. utilizando a notação de somatório.

1.7 Arredondamento e Truncagem

Arredondamento[4] e *truncagem* são métodos para escrever números com precisão delimitada.

Para *arredondar* um número para a k-ésima casa decimal, basta observar a k+1-ésima casa. Se a k+1-ésima casa decimal for 0, 1, 2, 3 ou 4, mantém-se a k-ésima casa decimal; se a k+1-ésima casa decimal for 5, 6, 7, 8 ou 9, soma-se 1 à k-ésima casa decimal. Como exercício, releia a frase anterior substituindo 'k-ésima' por 'primeira' e 'k+1-ésima' por 'segunda', aplicando esta regra para o número 153.654321. Note que **deve-se sempre avaliar o número original para realizar o arredondamento**. Arredondamentos são comuns, por exemplo, ao calcularmos um índice de preço ou um montante de pagamento sobre o qual incidiu certa taxa de juros.

Para *truncar* um número para a k-ésima casa decimal, basta eliminar a k+1-ésima casa decimal e suas subsequentes. Como exercício, releia a frase anterior substituindo 'k-ésima' por 'primeira' e 'k+1-ésima' por 'segunda', aplicando esta regra novamente para o número 153.654321. Compare com os valores arredondados e note que pode-se utilizar números já truncados para continuar a reduzir a precisão sem a necessidade de conhecer o valor original. Truncagens são comuns, por exemplo, para representar idades e ao calcular os graus G1 e G2 da PUCRS. Assim, se o cálculo do seu G1 resultar em 6.99999999, o sistema irá truncar para 6.9, e não arredondar para 7.0.

[4] Esta é a regra do *arredondamento para o número mais próximo*.

Exemplo 1.3. (Arredondamento e truncagem)

Decimais	Arredondamento	Truncagem
6	153.654321	153.654321
5	153.65432	153.65432
4	153.6543	153.6543
3	153.654	153.654
2	153.65	153.65
1	153.7	153.6
0	154	153
-1	150	150
-2	200	100

```
# Usando base R
options(digits = 10)            # Ajustando para apresentação de 10 dígitos (padrão: 7)
for(i in 6:-2){ print(round(153.654321, dig = i)) }        # 'digits' casas decimais
```

```
## [1] 153.654321
## [1] 153.65432
## [1] 153.6543
## [1] 153.654
## [1] 153.65
## [1] 153.7
## [1] 154
## [1] 150
## [1] 200
```

```
trunc <- function(x, ..., dig = 0) base::trunc(x*10^dig, ...)/10^dig   # Aprimorando
for(i in 6:-2){ print(trunc(153.654321, dig = i)) }        # Precisão de i decimais
```

```
## [1] 153.654321
## [1] 153.65432
## [1] 153.6543
## [1] 153.654
## [1] 153.65
## [1] 153.6
## [1] 153
## [1] 150
## [1] 100
```

```
# Usando o pacote plyr
plyr::round_any(153.654321, .01, round)          # dig = 2 em round
```

```
## [1] 153.65
```

```
plyr::round_any(153.654321, .0001, floor)          # dig = 4 em trunc
```

```
## [1] 153.6543
```

```
plyr::round_any(153.654321, 1, round)              # dig = 0 em round
```

```
## [1] 154
```

```
plyr::round_any(153.654321, 100, round)            # dig = -2 em round
```

```
## [1] 200
```

1.8 Outros símbolos e expressões

- $\sim$: tem distribuição.
- $\approx$: aproximadamente.
- $\#$: número de.
- $\pm/\mp$: mais ou menos/menos ou mais.
- $\triangle$: fim do Teorema.
- i.e.: *id est*, expressão em Latim que significa 'isto é'.
- e.g.: *exempli gratia*, expressão em Latim que significa 'por exemplo'.

1.8.1 Alfabeto grego

Maiúscula	Minúscula	Nome	Maiúscula	Minúscula	Nome
A	α	Alfa	N	ν	Nü
B	β	Beta	Ξ	ξ	Csi
Γ	γ	Gama	O	o	Ômicron
Δ	δ	Delta	Π	π, ϖ	Pi
E	ϵ, ε	Épsilon	P	ρ, ϱ	Rô
Z	ζ	Zeta	Σ	σ, ς	Sigma
H	η	Eta	T	τ	Tau
Θ	θ, ϑ	Teta	Υ	υ	Úpsilon
I	ι	Iota	Φ	ϕ, φ	Fi
K	κ, $\varkappa$	Capa	X	χ	Qui
Λ	λ	Lambda	Ψ	ψ	Psi
M	μ	Mü	Ω	ω	Ômega

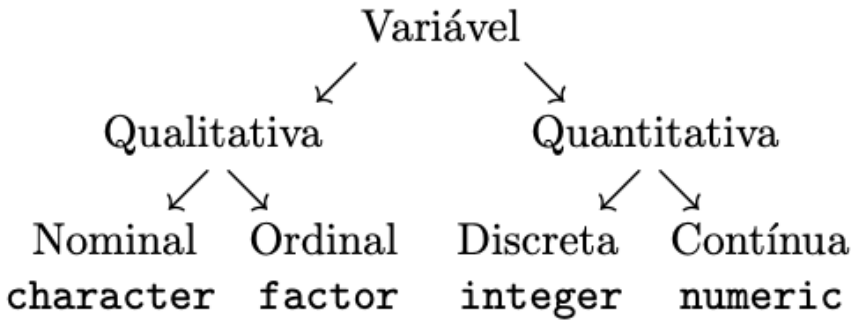
Chapter 2

Estatística Descritiva

A *Estatística Descritiva* está diretamente ligada à organização e descrição dos dados. É utilizada para avaliar como as observações se distribuem, onde estão posicionadas e como se apresentam em termos de dispersão e associação. Neste capítulo serão introduzidos conceitos e métodos descritivos, ponto de partida da *análise exploratória de dados*, passo fundamental para análises estatísticas mais avançadas.

2.1 Variáveis

Variável é uma característica medida nos universos ou amostras. As *variáveis qualitativas* ou *atributos* avaliam características não numéricas no conjunto de interesse, como gênereo, time de futebol e nível de escolaridade. As *variáveis quantitativas* medem características numéricas, como número de alunos prestando atenção ou tempo de uma música em segundos. Podem ser classificadas conforme a Figura a seguir.



2.1.1 Variável qualitativa nominal

Variáveis *qualitativas nominais* possuem o menor grau de informação dentre os quatro tipos propostos, permitindo apenas a avaliação de frequências e ordenações arbitrárias. Aplicam-se em avaliações de grupos não ordenados, tais como 'gênero', 'religião', 'raça', 'cor preferida', 'bairro onde reside', 'time de futebol do coração', etc.

Exemplo 2.1. (Time de futebol do coração) Suponha um lugar onde tudo seja tratado de maneira dicotômica[1]. Como exercício, no primeiro dia de aula de Estatística as pessoas são questionadas quanto ao 'time de futebol do coração' através do voto secreto em uma cédula, onde estão listados os dois times locais. Não existe informação prévia que obrigue a dispor na listagem qualquer time antes ou depois de outro. Por este motivo optou-se pela ordenação alfabética – apesar do princípio de tumulto –, resultando na lista

Maragato F.C.
Ximango F.C.

Os mais tradicionalistas gritavam palavras de ordem, preferindo a grafia

Chimango F.C.
Maragato F.C.

2.1.2 Variável qualitativa ordinal

Variáveis *qualitativas ordinais* possuem grau de informação maior em relação às nominais pois são dotadas de uma ordenação prévia, permitindo comparações entre as observações. As variáveis de natureza ordinal são utilizadas quando avaliam-se medidas tais como 'colocação em um torneio esportivo', 'grau de escolaridade', 'classificação de um restaurante quanto à qualidade da comida', etc.

Exemplo 2.2. (Colocação no vestibular) A variável 'colocação geral no vestibular' é classificada como qualitativa ordinal pois indica a ordenação do vestibulando em comparação aos demais, mesmo que não se conheça a nota

[1] *Dicotomia* é o ato de segmentar um conjunto em dois subconjuntos mutuamente excludentes, i.e., um elemento pode pertencer somente a um dos subconjuntos.

final de cada candidato.

Exemplo 2.3. (Escala de Likert) Quando deseja-se medir o grau de satisfação em relação a algum bem ou serviço, pode-se utilizar a Escala de Likert de k níveis. Uma vantagem de utilizar k par é que obriga-se o respondente a se posicionar a favor/contra, acima/abaixo. Se um empresário utilizar $k = 4$, pode fazer 1: Ruim, 2: Regular, 3: Bom, 4: Ótimo. Se $k = 5$, pode-se considerar 1: Péssimo, 2: Ruim, 3: Regular, 4: Bom, 5: Ótimo.

Exemplo 2.4. Corrida maluca) Suponha uma corrida disputada em Imaginationland[2], na qual Rubinho Barrichello tenha chegado na primeira colocação e Ayrton Senna na décima nona. As únicas informações de que dispomos apontam que i) Barrichello chegou antes de Senna, ii) ninguém chegou antes de Barrichello, iii) há 17 intermediários e iv) de fato, tudo aconteceu em Imaginationland.

2.1.3 Variável quantitativa discreta

Uma variável *quantitativa discreta* assume apenas valores inteiros, i.e., discretos. Tecnicamente as variáveis discretas são caracterizadas por conjuntos enumeráveis[3] finitos ou infinitos.

Exemplo 2.5. (Número de filhos) Suponha que deseja-se observar o número de filhos de mulheres atendidas em um hospital. Para cada mulher entrevistada, o conjunto de possíveis respostas para a pergunta 'quantos filhos a senhora tem?' é $F = \{0, 1, 2, \dots, k\}$, onde k é o número máximo de filhos que uma mulher possa ter ao longo de sua vida. O recorde mundial é $k = 69$, atribuído à russa Valentina Vassilyeva. Este é um conjunto enumerável finito.

Exemplo 2.6. (Pontos em um dado lançado k vezes) Suponha k lançamentos de um dado. Em cada lançamento é anotada a face resultante, somada aos valores obtidos nos $k - 1$ lançamentos anteriores. O conjunto de possíveis resultados deste experimento é $S = \{k, k + 1, \dots, 6k\}$. Este é um conjunto enumerável finito. Como exercício, faça $k = 4$ e releia a sentença anterior substituindo os valores.

Exemplo 2.7. (Consumo de uma engrenagem moto-contínua) Suponha uma engrenagem eterna, com consumo medido em PAB[4]. O conjunto do número

[2] http://www.imdb.com/title/tt0995577

[3] Um conjunto *enumerável* é aquele em que se pode listar e contar os elementos.

[4] Passos Até a Bufunfa.

possível de passos é $S = \{1, 2, ...\}$. Este é um conjunto enumerável infinito.

Exemplo 2.8. (Pilcher's Squad) Norman Pilcher foi o criador da Drug Squad, e ganhou notoriedade nos anos 60 por prender artistas como Mick Jagger e John Lennon. O conjunto de artistas que o Sargento Pilcher poderia prender é $A = \{a_1, a_2, ..., a_k\}$, onde k representa o número de artistas disponíveis para serem presos. Este é um conjunto enumerável finito.

2.1.4 Variável quantitativa contínua

A classe de variáveis *quantitativas contínuas* é caracterizada por permitir a observação de qualquer subconjunto dos números reais como resultado, i.e., permite resultados não inteiros. É utilizada para avaliar tempo, distâncias, áreas, volumes ou qualquer outra grandeza numérica de caráter não enumerável[5]. Tal como nas variáveis discretas, é possível avaliar relações matemáticas entre os valores observados.

Exemplo 2.9. (Proporção de bulímicas) Suponha que um grupo de pesquisadores está interessado em avaliar a 'proporção de mulheres bulímicas no Rio Grande do Sul'. Este valor está obrigatoriamente entre 0 e 1 (ou 0% e 100%), podendo ser representado pelo conjunto não enumerável $\Omega = \{b \in \mathbb{R} : 0 \leq b \leq 1\}$.

Exemplo 2.10. (Idade) A variável 'idade' é classificada como quantitativa contínua por representar uma noção temporal. Caso haja interesse, pode-se dizer que em certo instante do tempo João apresentou 31.990192013071629871269817323644 anos de idade. Na prática, porém, geralmente as idades são truncadas, sendo que João provavelmente afirmaria ter 31 anos de idade mesmo um dia antes do seu 32ºaniversário. Na melhor das situações as idades são observadas com precisão de dias, calculando-se a idade do indivíduo pela diferença entre o dia de hoje e o seu dia de nascimento, convertendo o valor para anos. O conjunto dos possíveis tempos de vida de um ser humano é dado por $\Omega = \{t \in \mathbb{R} : 0 < t \leq T\}$, onde T é a idade máxima em anos que um ser humano pode atingir. Segundo o *Guiness World Records*, $T = 122.44931506849315$, alcançado pela francesa Jeanne Louise Calment. Ω é dito não enumerável visto não ser possível contabilizar o seu número de elementos.

[5] Um conjunto *não enumerável* possui infinitos elementos, sendo impossível listá-los segundo alguma regra.

Exemplo 2.11. (Descendo o nível) Suponha que um grupo de pessoas foi avaliado em relação à variável 'idade' medida em anos, considerando-se a hora e minuto do nascimento. É possível transformá-la na variável 'idade discreta' simplesmente truncando os valores observados. Da mesma forma, pode-se tranformá-la na variável 'idade ordinal', classificando-a de acordo com a tabela a seguir.

i	Faixa etária	Classificação
1	Até 10 anos	Criança
2	$10 \vdash 13$	Pré-adolescente
3	$13 \vdash 18$	Adolescente
4	$18 \vdash 35$	Adulto jovem
5	$35 \vdash 45$	Adulto
6	$45 \vdash 65$	Adulto maduro
7	$65 \vdash 75$	Idoso jovem
8	$75 +$	Idoso

Note que se uma pessoa tem 31.990192013071629871269817323644 anos de idade (contínua), pode-se considerar a idade truncada de 31 anos (discreta) e classificá-la como um 'adulto jovem' (ordinal). Porém, dado que uma pessoa é classificada como adulto jovem, é possível apenas afirmar que ela tem idade entre 18 anos (completos) e 35 anos (incompletos) segundo a classificação proposta.

Cada tipo de variável apresenta um nível de informação que deve ser respeitado. É possível ir de um nível maior de classificação para um nível menor, mas jamais ao contrário. É válido lembrar que perde-se informação ao descer o nível de classificação da variável. É bastante comum, porém, encontrar trabalhos utilizando níveis de classificação inapropriados, conduzindo a técnicas não adequadas que implicam em conclusões equivocadas.

Exercício 2.1. Classifique as variáveis abaixo (qualitativa nominal/ordinal, quantitativa discreta/contínua).

a. Número de geladeiras em casa
b. Temperaturas da água da piscina em um dia de verão
c. Número de suicídios em uma cidade no decorrer do ano passado
d. Concentração de chumbo em uma amostra de água
e. Lista de editoras de livros
f. Grau de satisfação dos clientes que frequentam uma rinha de galo
g. Marcas de amaciantes para roupas
h. Tempo que um paciente sobrevive após determinado diagnóstico
i. Participação de mercado (*market share*)
j. Classificação em uma corrida de banheiras
k. Tempo final de cada corredor

l. Lista dos nomes das banheiras participantes, tal como 'Dick Vigarista' e 'Trollface'
m. Distância de Estambul ao Rio de Janeiro

2.2 Distribuição de Frequência

2.2.1 Dados brutos, Rol e Estatísticas de Ordem

Quando observa-se alguma variável de interesse, em geral anotam-se os resultados na ordem em que aparecem. Esta lista de dados não ordenada é conhecida como *lista de dados brutos*. Quando ordenam-se estes dados – em ordem crescente ou decrescente – obtém-se um *rol*, dando origem às *estatísticas de ordem*. Em uma distribuição de n elementos x_1, x_2, ..., x_n observados sequencialmente, denotam-se os dados ordenados de forma crescente por $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ e, analogamente, $x_{(n)}$, $x_{(n-1)}$, ..., $x_{(1)}$ para a ordenação decrescente.

Exemplo 2.12. (Rol) Se ordenarmos os dados observados da variável X:'número de passos até a lixiera mais próxima', obtemos o rol conforme tabela a seguir. O menor número de passos caminhados foi sete, representado por $x_{(1)} = 7$, e o maior foi quatrocentos e dois, representado por $x_{(6)} = 402$.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$
7	20	124	186	191	402

```
(x <- c(186,402,191,20,7,124))  # Criando e apresentando o vetor original de dados bru
```

```
## [1] 186 402 191  20   7 124
```

```
sort(x)                         # Apresentando o rol, ou vetor ordenado. Teste ?order
```

```
## [1]   7  20 124 186 191 402
```

```
sort(x, decreasing = T)         # Ordem decrescente, onde T indica TRUE (padrão: FALSE
```

```
## [1] 402 191 186 124  20   7
```

Em um primeiro momento estas definições podem parecer ultrapassadas, mas são de grande importância na construção de métodos avançados de análise de dados. Como atualmente trabalham-se com bases de dados em formato eletrônico, em geral é fácil realizar a ordenação de grandes volumes de dados. É importante ressaltar, porém, que em certos casos é necessário muito poder de processamento para executar tais ordenações, podendo se tornarem impraticáveis pelo alto custo computacional.

Exercício 2.2. Considere o conjunto de dados $10, -4, 5, 7, 1, 3, 9$.

a. Obtenha o rol.
b. Indique e interprete $x_{(4)}$.

Exercício 2.3. Utilizando a função `sort`, encontre o rol das colunas `filhos` e `altura` disponíveis em http://filipezabala.com/data/hospital.txt.

2.2.2 Tabela de frequência univariada discreta

Listas muito longas, ainda que ordenadas, não costumam ser de fácil compreensão. Assim, a *tabela de frequência univariada discreta* é uma boa maneira de consolidar os dados de uma variável que assuma, como regra-de-bolso, até 10 diferentes valores. Esta tabela deve apresentar pelo menos uma coluna descrevendo a variável de interesse e uma coluna com a *frequência (da classe)*, i.e., o número de observações contempladas em cada categoria. Sugere-se também a apresentação de uma coluna indicando a classe, denotada por i conforme Tabela a seguir.

i	x_i	f_i	f_{r_i}	F_i	F_{r_i}	$_i$	$_{r_i}$
1	x_1	f_1	f_1/n	$F_1 = f_1$	F_1/n	$_1 = \ _2 + f_1 = n$	$_1/n = 1$
2	x_2	f_2	f_2/n	$F_2 = F_1 + f_2$	F_2/n	$_2 = \ _3 + f_2$	$_2/n$
3	x_3	f_3	f_3/n	$F_3 = F_2 + f_3$	F_3/n	$_3 = \ _4 + f_3$	$_3/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k-2$	x_{k-2}	f_{k-2}	f_{k-2}/n	$F_{k-2} = F_{k-3} + f_{k-2}$	F_{k-2}/n	$_{k-2} = \ _{k-1} + f_{k-2}$	$_{k-2}/n$
$k-1$	x_{k-1}	f_{k-1}	f_{k-1}/n	$F_{k-1} = F_{k-2} + f_{k-1}$	F_{k-1}/n	$_{k-1} = \ _k + f_{k-1}$	$_{k-1}/n$
k	x_k	f_k	f_k/n	$F_k = F_{k-1} + f_k = n$	$F_k/n = 1$	$_k = f_k$	$_k/n$
Total	-	n	1	-	-	-	-

Para a classe genérica i são calculadas as seguintes frequências:
- f_i: Frequência (simples/absoluta)
- f_{r_i}: Frequência relativa
- F_i: Frequência acumulada
- F_{r_i}: Frequência acumulada relativa
- $_i$: Frequência acumulada inversa
- $_{r_i}$: Frequência acumulada inversa relativa

Exemplo 2.13. (Número de filhos revisitado) Do Exemplo 2.5 observou-se a variável

X: 'número de filhos de mulheres atendidas em um hospital de Porto Alegre em 2019'.

A Tabela a seguir apresenta os dados na ordem em que foram observados. Este tipo de apresentação é bastante completo, mas dificulta a extração de informações relevantes. Como exercício, indique o número máximo de filhos observados na amostra.

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	2	21	2	41	1	61	3	81	0
2	0	22	3	42	1	62	0	82	1
3	1	23	1	43	4	63	2	83	2
4	2	24	2	44	1	64	0	84	2
5	4	25	2	45	1	65	2	85	2
6	2	26	1	46	3	66	2	86	2
7	1	27	4	47	1	67	2	87	2
8	4	28	0	48	1	68	1	88	4
9	2	29	1	49	4	69	2	89	0
10	3	30	6	50	2	70	3	90	2
11	3	31	1	51	2	71	1	91	1
12	2	32	1	52	4	72	3	92	3
13	3	33	1	53	1	73	1	93	3
14	2	34	1	54	3	74	3	94	4
15	1	35	0	55	1	75	3	95	5
16	4	36	2	56	2	76	4	96	1
17	2	37	3	57	0	77	2	97	0
18	0	38	3	58	2	78	1	98	0
19	1	39	1	59	3	79	2	99	3
20	4	40	2	60	3	80	3	100	2

A Tabela a seguir apresenta o número de filhos ordenados, fornecendo ainda algumas frequências que auxiliam o entendimento da distribuição. Com a apresentação neste formato, facilmente observa-se o máximo de 6 filhos na amostra, ao contrário da tabela de dados brutos. Perde-se apenas a ordem na qual os dados foram observados, o que em geral não é do interesse do pesquisador.

i	x_i	f_i	f_{r_i}	F_i	F_{r_i}	$_i$	$_{r_i}$
1	0	11	$11/100 = 0.11$	11	$11/100 = 0.11$	$89 + 11 = 100$	$100/100 = 1$
2	1	27	$27/100 = 0.27$	$11 + 27 = 38$	$38/100 = 0.38$	$62 + 27 = 89$	$89/100 = 0.89$
3	2	30	$30/100 = 0.30$	$38 + 30 = 68$	$68/100 = 0.68$	$32 + 30 = 62$	$62/100 = 0.62$
4	3	19	$19/100 = 0.19$	$68 + 19 = 87$	$87/100 = 0.87$	$13 + 19 = 32$	$32/100 = 0.32$
5	4	11	$11/100 = 0.11$	$87 + 11 = 98$	$98/100 = 0.98$	$2 + 11 = 13$	$13/100 = 0.13$
6	5	1	$1/100 = 0.01$	$98 + 1 = 99$	$99/100 = 0.99$	$1 + 1 = 2$	$2/100 = 0.02$
7	6	1	$1/100 = 0.01$	$99 + 1 = 100$	$100/100 = 1$	1	$1/100 = 0.01$

i	x_i	f_i	f_{r_i}	F_i	F_{r_i}	$_i$	$_{r_i}$
Total	-	100	1	-	-	-	-

Note que a coluna i da Tabela acima indica a ordem da mulher entrevistada, enquanto na Tabela de dados brutos i indica a classe. Por exemplo, $i = 4$ indica a quarta mulher entrevistada, que no caso informou ter $x_4 = 2$ filhos. Na Tabela acima, $i = 4$ indica a quarta classe onde $x_4 = 3$, i.e., a classe das mulheres que possuem 3 filhos.

As únicas colunas que exigem a leitura dos dados brutos são a da variável x_i e a da frequência f_i; as demais são calculadas a partir de f_i. A seguir estão alguns exemplos de interpretação das frequências apresentadas na Tabela acima.
- $f_5 = 11$, i.e., 11 mulheres possuem 4 filhos
- $f_{r_5} = 0.11 = 11\%$, i.e., 11% das mulheres possuem 4 filhos
- $F_4 = 87$, i.e., 87 mulheres possuem até 3 filhos (ou 'de zero a 3 filhos', mas esta alternativa é menos elegante)
- $F_{r_3} = 0.68 = 68\%$, i.e., 68% das mulheres possuem até 2 filhos
- $_3 = 62$, i.e., 62 mulheres têm pelo menos 2 filhos
- $_{r_2} = 0.89 = 89\%$, i.e., 89% das mulheres têm pelo menos 1 filho

Exemplo 2.14. (Número de filhos R-visitado) Exemplo 2.13 utilizando R/RStudio.

```
hosp <- read.table('http://www.filipezabala.com/data/hospital.txt', head = T)
dim(hosp)       # Dimensão: 100 linhas por 2 colunas
```

```
## [1] 100   2
```

```
head(hosp)      # Apresenta as 6 primeiras linhas do objeto 'hosp'; teste tail(hosp, 10)
```

```
##   filhos altura
## 1      2   1.59
## 2      0   1.58
## 3      1   1.70
## 4      2   1.62
## 5      4   1.67
## 6      2   1.62
```

```
attach(hosp, warn=F)                    # Para deixar as colunas de 'hosp' disponíveis
(tab <- table(filhos))                  # Frequência (simples/absoluta)
```

```
## filhos
##  0  1  2  3  4  5  6
## 11 27 30 19 11  1  1
```

```r
prop.table(tab)                                  # Frequência relativa
```

```
## filhos
##    0    1    2    3    4    5    6
## 0.11 0.27 0.30 0.19 0.11 0.01 0.01
```

```r
cumsum(tab)                                      # Frequência acumulada
```

```
##   0   1   2   3   4   5   6
##  11  38  68  87  98  99 100
```

```r
round(cumsum(tab)/length(filhos),2)              # Frequência acumulada relativa
```

```
##    0    1    2    3    4    5    6
## 0.11 0.38 0.68 0.87 0.98 0.99 1.00
```

```r
cumsum(rev(tab))                                 # Frequência acumulada inversa
```

```
##   6   5   4   3   2   1   0
##   1   2  13  32  62  89 100
```

```r
round(cumsum(rev(tab))/length(filhos),2)  # Frequência acumulada inversa relativa
```

```
##    6    5    4    3    2    1    0
## 0.01 0.02 0.13 0.32 0.62 0.89 1.00
```

Exercício 2.4. Em uma fábrica retirou-se uma amostra de 50 peças de um lote de certo material e contou-se o número de defeitos em cada peça, apresentados na tabela a seguir.

i	# defeitos	f_i	fr_i	F_i	Fr_i	$_i$	$_{r_i}$
1	0	17					
2	1	10					
3	2						
4	3	8					
5	4	5					
6	5	1					
Total	-	50					

a. Classifique a variável 'número de defeitos'.

b. Qual a frequência absoluta da classe 3? Interprete.

c. Qual a frequência relativa da classe 3? Interprete.

d. Qual a frequência acumulada da classe 4? Interprete.

e. Qual a frequência acumulada relativa da classe 5? Interprete.

f. Represente os dados utilizando o gráfico que você considerar mais adequado.

2.2.3 Tabela de frequência univariada contínua

Quando uma variável assume mais de 10 diferentes valores, recomenda-se utilizar a *tabela de frequência univariada contínua*. A diferença para a tabela discreta da Seção 2.2.2 é que na contínua distribuem-se os valores em *intervalos de classe*, i.e., faixas de valores com certa amplitude. A principal vantagem desta abordagem é a capacidade de apresentar os dados de maneira enxuta. O contraponto, como em qualquer resumo de dados, é a perda da informação original.

Amplitude (h) e quantidade (k) de classes

Quando deseja-se apresentar a variável em intervalos de classe, é necessário determinar a *amplitude do intervalo de classe* (h) e a *quantidade de classes* (k) em que serão dispostos os dados. Apresentam-se a seguir três das principais regras para determinar h e k.

1. Sturges (Sturges, 1926) sugere que a amplitude do intervalo de classe seja calculada por

$$h_{St} = \frac{A}{k_{St}} = \frac{\max X - \min X}{1 + \log_2 n} \approx \frac{\max X - \min X}{1 + 3.322 \log_{10} n}, \tag{2.1}$$

onde A é a *amplitude (dos dados)* descrita na Seção 2.4.1, não devendo ser confundida com h. O denominador é obtido a partir da *expansão binomial*, na forma

$$n = \sum_{i=0}^{k-1} \binom{k-1}{i} = (1+1)^{k-1} = 2^{k-1}. \tag{2.2}$$

Da Equação (2.2) pode-se obter

$$k_{St} = \lceil 1 + \log_2 n \rceil = \lceil 1 + 3.322 \log_{10} n \rceil, \tag{2.3}$$

onde $\lceil\ \rceil$ indica a função *teto*, o menor inteiro consecutivo denotado por $\lceil x \rceil = min\{n \in \mathbb{N} | n \geq x\}$. Alguns pacotes computacionais atribuem o número de classes aplicando regras que encontrem um valor 'bonito' para a divisão. Tais valores são obtidos computando números que sejam 0, 1, 2 ou 5 vezes uma potência de 10, i.e., $a \times 10^b$, $a \in \{0, 1, 2, 5\}$ e $b \in \mathbb{N} \cup \{-1\}$.

Exemplo 2.15. (Sturges) Se forem observados $n = 100$ valores com amplitude (dos dados) $A = 0.23$, a amplitude da classe sugerida por Sturges é

$$h_{St} = \frac{0.23}{1 + \log_2 100} = 0.02875,$$

e a quantidade de classes

$$k_{St} = \lceil 1 + \log_2 100 \rceil = \lceil 7.644 \rceil = 8.$$

```
n <- length(altura)                # n=100, número de dados a serem tabulados
A <- diff(range(altura))           # Amplitude (dos dados, não da classe!)
ceiling(1 + log2(n))               # Pela Equação (2.3), usando log2
```

```
## [1] 8
```

```
ceiling(1 + 3.322*log10(n))        # Pela Equação (2.3), usando log10
```

```
## [1] 8
```

```
(kSt <- nclass.Sturges(altura))    # Pela função 'nclass.Sturges'
```

```
## [1] 8
```

```
(hSt <- A/kSt)                     # Pela Equação (2.1)
```

```
## [1] 0.02875
```

```
pretty(kSt)                        # Valores 'bonitos', (a=5, b=0) e (a=1, b=1)
```

```
## [1]  5 10
```

2. Scott (Scott, 1979) incorpora s, o desvio padrão amostral ao cálculo da amplitude do intervalo, na forma

$$h_{Sc} = \frac{3.5s}{n^{1/3}}. \tag{2.4}$$

O número de classes de Scott pode ser obtido por

$$k_{Sc} = \left\lceil \frac{A}{h_{Sc}} \right\rceil = \left\lceil \frac{maxX - minX}{3.5sn^{-1/3}} \right\rceil. \tag{2.5}$$

Exemplo 2.16. (Scott) Se forem observados $n = 100$ valores com desvio padrão amostral $s = 0.045268559$, a amplitude da classe sugerida por Scott é

$$h_{Sc} = \frac{3.5 \times 0.045268559}{100^{1/3}} = 0.034134854.$$

Se $A = 0.23$, a quantidade de classes é

$$k_{Sc} = \left\lceil \frac{0.23}{0.034134854} \right\rceil = \lceil 6.7379811 \rceil = 7.$$

```
n <- length(altura)                  # n=100, número de observações a serem tabuladas
s <- sd(altura)                      # s=0.045268559, desvio padrão amostral
A <- diff(range(altura))             # Amplitude (dos dados, não da classe!)
(hSc <- 3.5*s/n^(1/3))               # Pela Equação (2.4)
```

```
## [1] 0.03413485378
```

```
ceiling(A/hSc)                       # k sugerido por Scott, Equação (2.5)
```

```
## [1] 7
```

```
(kSc <- nclass.scott(altura))        # k obtido pela função 'nclass.scott'
```

```
## [1] 7
```

```
pretty(kSc)                          # Valores 'bonitos', (a=5, b=0) e (a=1, b=1)
```

```
## [1]  5 10
```

3. Freedman-Diaconis (Freedman and Diaconis, 1981) inserem a *amplitude interquartílica* no cálculo da amplitude do intervalo, na forma

$$h_{FD} = 2\frac{AI}{n^{1/3}}, \tag{2.6}$$

onde $AI = Q_3 - Q_1$ é a amplitude interquartílica. O número de classes obtido como consequência da aplicação da relação de Freedman-Diaconis é

$$k_{FD} = \left\lceil \frac{A}{h_{FD}} \right\rceil = \left\lceil \frac{maxX - minX}{2\,AI\,n^{-1/3}} \right\rceil. \tag{2.7}$$

Exemplo 2.17. (Freedman-Diaconis) Se forem observados $n = 100$ valores com amplitude interquartílica de $AI = 0.0525$, a amplitude da classe sugerida por Freedman-Diaconis é

$$h_{FD} = \frac{2 \times 0.0525}{100^{1/3}} = 0.022621564.$$

Se $A = 0.23$, e a quantidade de classes

$$k_{FD} = \left\lceil \frac{0.23}{0.022621564} \right\rceil = \lceil 10.16729 \rceil = 11.$$

```
n <- length(altura)                  # n=100, número de observações a serem tabuladas
(Q <- quantile(altura, c(1/4,3/4)))  # Primeiro e terceiro quartis
```

```
##     25%     75%
## 1.5975 1.6500
```

```
(AI <- diff(as.numeric(Q)))      # Amplitude Interquatílica
```

```
## [1] 0.0525
```

```
(hFD <- 2*AI/n^(1/3))            # Pela Equação (2.6)
```

```
## [1] 0.02262156425
```

```
A <- diff(range(altura))         # Amplitude (dos dados, não da classe ou interquartíli
ceiling(A/hFD)                   # k sugerido por Freedman-Diaconis, Equação (2.7)
```

```
## [1] 11
```

```
(kFD <- nclass.FD(altura))       # Pela função 'nclass.FD'
```

```
## [1] 11
```

```
pretty(kFD)                      # Valores 'bonitos', (a=1, b=1) e (a=2, b=1)
```

```
## [1] 10 20
```

(Hyndman, 1995) argumenta que as regras de Scott e Freedman-Diaconis são tão simples quanto a regra de Sturges, mas melhor fundamentadas na teoria estatística. Além disso, a regra de Sturges funciona bem para tamanhos de amostra moderados ($n < 200$), mas não para valores grandes de n.

Exemplo 2.18. (Comparando os três métodos) Foi realizada uma simulação com tamanhos de amostra $n = 10^i$, $i \in \{1, 2, \ldots, 6\}$, indicando o número de classes sugerido por cada método.

```
NC <- function(x) c(i = i, n = 10^i,                # Quantidades simuladas
                    Sturges = nclass.Sturges(x),     # Sturges (1926)
                    Scott = nclass.scott(x),         # Scott (1979)
                    FD = nclass.FD(x))               # Freedman-Diaconis (1981)
for(i in 1:6){set.seed(i); print(NC(rnorm(10^i)))}  # Pode ser demorado para i>6
```

```
##       i       n Sturges   Scott      FD
##       1      10       5       2       3
##       i       n Sturges   Scott      FD
##       2     100       8       6       7
##       i       n Sturges   Scott      FD
##       3    1000      11      19      25
##       i       n Sturges   Scott      FD
##       4   10000      15      44      56
##       i       n Sturges   Scott      FD
##       5  100000      18     112     145
##       i       n Sturges   Scott      FD
##       6 1000000      21     278     360
```

Exemplo 2.19. (Alturas de mulheres) Seja a variável

> Y: 'altura de mulheres atendidas em um hospital de Porto Alegre em 2019'.

A Tabela abaixo apresenta os dados brutos. Este tipo de apresentação é bastante completo, mas dificulta a extração de informações relevantes. Como exercício, indique quantas mulheres têm altura entre 1.70m e 1.75m a partir desta tabela.

i	y_i	i	y_i	i	y_i	i	y_i
1	1.59	26	1.61	51	1.64	76	1.62
2	1.58	27	1.61	52	1.57	77	1.54
3	1.70	28	1.60	53	1.65	78	1.64
4	1.62	29	1.61	54	1.69	79	1.66
5	1.67	30	1.64	55	1.65	80	1.56
6	1.62	31	1.59	56	1.62	81	1.64
7	1.69	32	1.60	57	1.68	82	1.60
8	1.60	33	1.62	58	1.60	83	1.68
9	1.61	34	1.53	59	1.68	84	1.65
10	1.58	35	1.58	60	1.59	85	1.65
11	1.64	36	1.60	61	1.70	86	1.64
12	1.72	37	1.61	62	1.65	87	1.55
13	1.74	38	1.67	63	1.51	88	1.66
14	1.63	39	1.68	64	1.66	89	1.59
15	1.64	40	1.56	65	1.52	90	1.66
16	1.63	41	1.58	66	1.60	91	1.69
17	1.59	42	1.66	67	1.62	92	1.61
18	1.64	43	1.59	68	1.68	93	1.58
19	1.59	44	1.67	69	1.65	94	1.73
20	1.65	45	1.62	70	1.61	95	1.56
21	1.63	46	1.55	71	1.56	96	1.59
22	1.64	47	1.64	72	1.65	97	1.65
23	1.64	48	1.62	73	1.62	98	1.63
24	1.62	49	1.65	74	1.63	99	1.70
25	1.66	50	1.66	75	1.57	100	1.60

Para colocar estes valores em uma tabela de frequência, obteve-se $k_{St} = 8$ pela regra de Sturges, e pelo resultado de `pretty(8)` decidiu-se por 5 classes. Como exercício, obtenha k_{Sc} e k_{FD}.

A Tabela abaixo apresenta as alturas agrupadas em cinco classes de amplitude 5cm, fornecendo ainda algumas frequências que auxiliam o entendimento da distribuição. Facilmente observam-se 6 mulheres com altura entre 1.70m e 1.75m,[6]

[6]Note que a simbologia $1.70 \vdash 1.75$ indica a inclusão de 1.70 e a exclusão de 1.75, i.e., este é um intervalo fechado à esquerda e aberto à direita. Equivale às notações $[1.70, 1.75[$ (mais moderna) ou $[1.70, 1.75)$ (mais antiga).

ao contrário da tabela de dados brutos. Note, porém, que não é possível saber exatamente a altura de cada uma destas 6 mulheres. Isso acontece pois resumo implica em perda de informação, cabendo ao pesquisador decidir quando e como resumir os dados.

i	y_i	f_i	f_{r_i}	F_i	F_{r_i}	$_i$	$_{r_i}$
1	$1.50 \vdash 1.55$	4	0.04	4	0.04	$96 + 4 = 100$	$100/100 = 1$
2	$1.55 \vdash 1.60$	21	0.21	$4 + 21 = 25$	0.25	$75 + 21 = 96$	$96/100 = 0.96$
3	$1.60 \vdash 1.65$	41	0.41	$25 + 41 = 66$	0.66	$34 + 41 = 75$	$75/100 = 0.75$
4	$1.65 \vdash 1.70$	28	0.28	$66 + 28 = 94$	0.94	$6 + 28 = 34$	$34/100 = 0.34$
5	$1.70 \vdash 1.75$	6	0.06	$94 + 6 = 100$	1	6	$6/100 = 0.06$
Total	-	100	1	-	-	-	-

A seguir estão alguns exemplos de interpretação das frequências apresentadas na Tabela acima.
- $f_5 = 6$, i.e., 6 mulheres têm entre 1.70m e 1.75m de altura
- $f_{r_5} = 0.06 = 6\%$, i.e., 6% das mulheres tem entre 1.70m e 1.75m de altura
- $F_4 = 94$, i.e., 94 mulheres têm até 1.70m de altura, ou de 1.50m a 1.70m
- $F_{r_2} = 0.25 = 25\%$, i.e., 25% das mulheres tem até 1.60m de altura, ou de 1.50m a 1.60m
- $_3 = 75$, i.e., 75 mulheres têm pelo menos 1.60m de altura
- $_{r_4} = 0.34 = 34\%$, i.e., 34% das mulheres tem pelo menos 1.65m de altura

Exemplo 2.20. (Alturas de mulheres R-visitado) Exemplo 2.19 utilizando R/RStudio.

```
hosp <- read.table('http://www.filipezabala.com/data/hospital.txt', head = T)
dim(hosp)       # Dimensão: 100 linhas por 2 colunas
```

```
## [1] 100   2
```

```
head(hosp)      # Apresenta as 6 primeiras linhas do objeto 'hosp'; teste tail(hosp, 1
```

```
##   filhos altura
## 1      2   1.59
## 2      0   1.58
## 3      1   1.70
## 4      2   1.62
## 5      4   1.67
## 6      2   1.62
```

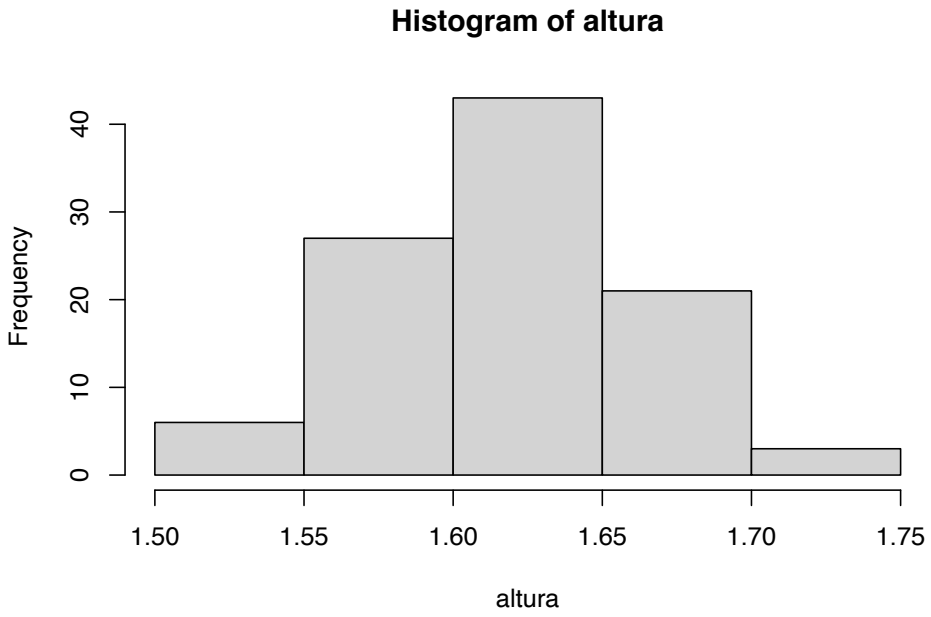
```
attach(hosp, warn=F)                # Para deixar as colunas de 'hosp' disponíveis
pretty(nclass.Sturges(altura))      # Valores 'bonitos' para o número de classes
```

```
## [1]  5 10
```

```r
hist(altura)$breaks                    # Quebras de valores gerados com a função 'hist'
```

```
## [1] 1.50 1.55 1.60 1.65 1.70 1.75
```

```r
(f <- hist(altura)$counts)             # Frequências das classes
```




```
## [1]  6 27 43 21  3
```

```r
cumsum(f)                              # Frequência acumulada
```

```
## [1]   6  33  76  97 100
```

```r
round(cumsum(f)/length(altura),2)      # Frequência acumulada relativa
```

```
## [1] 0.06 0.33 0.76 0.97 1.00
```

```r
cumsum(rev(f))                         # Frequência acumulada inversa
```

```
## [1]   3  24  67  94 100
```

```r
round(cumsum(rev(f))/length(altura),2) # Frequência acumulada inversa relativa
```

```
## [1] 0.03 0.24 0.67 0.94 1.00
```

2.3 Medidas de Posição (ou Localização)

2.3.1 Mínimo e Máximo

O *mínimo* de uma distribuição é o menor valor observado desta distribuição; de forma análoga, o *máximo* é o maior valor. São estatísticas de ordem, mais especificamente os extremos de um conjunto de dados ordenado (rol). Para uma distribuição de n elementos são denotadas por $\min X = x_{(1)}$ e $\max X = x_{(n)}$.

Apesar da simplicidade destas medidas, existem considerações teóricas sofisticadas a seu respeito. Para maiores detalhes, vide (Kotz and Nadarajah, 2000).

Exemplo 2.21. (Mínimo e máximo) Suponha novamente as $n = 100$ observações da variável Y: 'altura de mulheres atendidas em um certo hospital público de Porto Alegre em 2019', apresentadas no Exemplo 2.19. O mínimo e o máximo são denotados, respectivamente, por $\min Y = y_{(1)} = 1.51$ e $\max Y = y_{(100)} = 1.74$.

```
attach(read.table('http://www.filipezabala.com/data/hospital.txt', head=T), warn=F)
min(altura)     # Mínimo
```

```
## [1] 1.51
```

```
max(altura)     # Máximo
```

```
## [1] 1.74
```

```
range(altura)   # A função 'range' fornece o mínimo e o máximo
```

```
## [1] 1.51 1.74
```

2.3.2 Média (Aritmética Simples)

A *média (aritmética simples)* ou *valor esperado* é uma das medidas mais importantes da Estatística devido às suas propriedades e relativa facilidade de cálculo. A média da variável X é simbolizada genericamente por μ quando refere-se à média universal, e por $\bar{x}_n$ quando refere-se à média amostral. Pode-se utilizar a notação $\bar{x}_n$ para indicar o tamanho da amostra. Suas expressões no universo a na amostra são dadas respectivamente pelas equações (2.8) e (2.9). Por distribuir a soma dos valores da distribuição pelo número de observações, a média é uma medida que indica centro de massa.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{2.8}$$

$$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.9}$$

Exemplo 2.22. (Média aritmética simples) Suponha novamente os dados do Exemplo 1.2. O número médio de passos até a lixeira mais próxima foi de

$$\bar{x}_6 = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{186 + 402 + 191 + 20 + 7 + 124}{6} = \frac{930}{6} = 155.$$

```
x <- c(186,402,191,20,7,124)      # Vetor de dados brutos
mean(x)                           # Aplica as Equações (2.8) e (2.9). Veja ?mean
```

```
## [1] 155
```

2.3.3 Total

Total é a soma de todos os valores de uma variável. É expresso pelas equações (2.10) e (2.11).

$$\tau = \sum_{i=1}^{N} x_i \tag{2.10}$$

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} x_i = N\bar{x}_n, \tag{2.11}$$

onde $\bar{x}_n$ é a *média amostral*, apresentada na Equação (2.9).

Exemplo 2.23. (Total) Suponha novamente os dados do Exemplo @ref{exm:media-a-s}. Se alguém precisar de uma lixeira 20 vezes na capital gaúcha, estima-se que o número total de passos a serem caminhados é de

$$\hat{\tau} = \frac{20}{6} \times 930 = 20 \times 155 = 3100.$$

```
N <- 20                           # Tamanho do universo
x <- c(186,402,191,20,7,124)      # Vetor de dados brutos
N*mean(x)                         # Equação (2.11)
```

```
## [1] 3100
```

2.3.4 Média Quadrática

A *média quadrática* é a média dos valores ao quadrado, utilizada no cálculo das variâncias. É expressa por

$$Q^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}. \tag{2.12}$$

O *valor quadrático médio* é a raiz quadrada da média quadrática, denotado por

$$Q = \sqrt{Q^2}. \tag{2.13}$$

Exemplo 2.24. (Média quadrática e valor quadrático médio) A média quadrática dos valores 186, 402, 191, 20, 7 e 124 é

$$Q^2 = \frac{\sum_{i=1}^{6} x_i^2}{6} = \frac{186^2 + 402^2 + 191^2 + 20^2 + 7^2 + 124^2}{6} = \frac{248506}{6} = 41417.\bar{6}.$$

O valor quadrático médio destes valores é

$$Q = \sqrt{41417.\bar{6}} \approx 203.5133.$$

```
x <- c(186,402,191,20,7,124)     # Vetor de dados brutos
(mq <- mean(x^2))                # Eq. (2.12), compare com mean(x)^2
```

```
## [1] 41417.66667
```

```
sqrt(mq)                         # Eq. (2.13), valor quadrático médio
```

```
## [1] 203.5133083
```

2.3.5 Moda

A(s) *moda(s)* é (são) o(s) valor(es) mais frequente(s) de uma distribuição. Quando existe apenas uma moda, a distribuição é conhecida como *unimodal*. Se existirem duas modas, a distribuição é *bimodal*. Três modas configuram uma distribuição *trimodal*, e quatro ou mais modas indicam uma distribuição *multimodal*. Distribuições com frequências equivalentes para todos os valores são ditas *amodais*. Quando os dados estão agrupados deve-se indicar a *classe modal*, i.e., a classe de maior frequência. O esforço computacional para calcular a moda é realizar uma contagem.
No R existe a função `Mode` do pacote `pracma`, mas ela só funciona bem no caso unimodal. Por isso a seguir está apresentada a função `Modes`, adaptada da sugestão de digEmAll nesta discussão do StackOverflow. Nos exemplos a seguir são comparadas as duas abordagens.

```
# Função Modes
Modes <- function(x) {
  ux <- sort(unique(x))
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}
```

Exemplo 2.25. (Unimodal) A moda do conjunto de dados 4, 7, 1, 3, 3, 9 é $Mo = 3$, pois ele apresenta frequência 2 enquanto os demais valores têm frequência 1. Esta é uma distribuição unimodal.

```
dat <- c(4,7,1,3,3,9)
Modes(dat)
```

```
## [1] 3
```

```
pracma::Mode(dat)
```

```
## [1] 3
```

Exemplo 2.26. (Bimodal) As modas do conjunto de dados 4, 7, 1, 3, 3, 9, 7 são $Mo' = 3$ e $Mo'' = 7$, pois ambos têm frequência 2 enquanto os demais valores têm frequência 1. A ordem de apresentação é indiferente. Esta é uma distribuição bimodal.

```
dat <- c(4,7,1,3,3,9,7)
Modes(dat)
```

```
## [1] 3 7
```

```
pracma::Mode(dat)
```

```
## [1] 3
```

Exemplo 2.27. (Amodal) O conjunto de dados 4, 7, 1, 3, 9 é dito *amodal* pois todos os valores têm frequência 1.

```
dat <- c(4,7,1,3,9)
Modes(dat)    # se todos são moda, ninguém é moda
```

```
## [1] 1 3 4 7 9
```

```
pracma::Mode(dat)
```

```
## [1] 1
```

Exemplo 2.28. (Moda para dados agrupados) No Exemplo 2.19 observa-se que $f_3 = 41$ é a maior frequência. A classe modal é portanto a terceira, compreendida entre os valores 1.60 e 1.65.

2.3.6 Separatrizes

Separatrizes ou quantis[7] são medidas que dividem um conjunto de dados ordenados em k partes iguais. O método básico consiste em obter um rol dos dados e encontrar (ainda que de forma aproximada) os valores que repartem a distribuição de acordo com o k desejado. O esforço computacional para calcular quaisquer separatrizes é, portanto, o de realizar a ordenação dos dados.

[7] Pronuncia-se *quantís*.

Mediana ($k = 2$)

A *mediana* é a medida que divide metade dos dados ordenados (rol) à sua esquerda e a outra metade à sua direita, i.e., é a medida central em termos de ordenação. Sua posição é a média entre a primeira e última posições, dada por

$$Pos = \frac{1+n}{2} \tag{2.14}$$

Exemplo 2.29. (Mediana para n ímpar) Seja o conjunto de dados 10, -4, 11, 12, 1, 5, 15, formado por $n = 7$ valores. Quando ordenado obtemos o rol -4, 1, 5, 10, 11, 12, 15. Considerando $k = 2$, obtém-se a separatriz $Md = 10$, pois ela divide o conjunto em duas partes de mesmo tamanho (três valores abaixo da mediana 10 e três valores acima). Sua posição é dada por $Pos = \frac{1+7}{2} = 4$.

```
x <- c(10, -4, 11, 12, 1, 5, 15)
(n <- length(x))
```

```
## [1] 7
```

```
(pos <- (n+1)/2)
```

```
## [1] 4
```

```
sort(x)
```

```
## [1] -4  1  5 10 11 12 15
```

```
median(x)
```

```
## [1] 10
```

Quando o número de observações é par, basta tomar a média dos dois valores centrais do rol.

Exemplo 2.30. (Mediana para n par) Seja o conjunto de dados 15, -4, 11, 12, 1, 5, formado por $n = 6$ valores. Quando ordenado obtemos o rol -4, 1, 5, 11, 12, 15. Considerando novamente $k = 2$, obtém-se a separatriz $Md = \frac{5+11}{2} = 8$, pois ela divide o conjunto em duas partes de mesmo tamanho (três valores abaixo de 8 e três valores acima). Sua posição é dada por $Pos = \frac{1+6}{2} = 3.5$, i.e., a mediana é um valor intermediário entre a terceira e quarta posições.

```
x <- c(15, -4, 11, 12, 1, 5)
(n <- length(x))
```

```
## [1] 6
```

```
(pos <- (n+1)/2)
```

```
## [1] 3.5
```

```
sort(x)
```

```
## [1] -4  1  5 11 12 15
```

```
median(x)
```

```
## [1] 8
```

Separatrizes

Pode-se dividir um conjunto de dados em k setores, sendo os principais apresentados na tabela a seguir

k	Nome	Simbologia
2	Mediana	Md
3	Tercil	T_1, T_2
4	Quartil	Q_1, Q_2, Q_3
10	Decil	D_1, D_2, ..., D_9
100	Percentil	P_1, P_2, ..., P_{99}

Exemplo 2.31. (Separatrizes) A função `quantile` apresenta nove métodos para obtenção de separatrizes, portanto recomenda-se a leitura da documentação para maiores detalhes. Com ela pode-se facilmente obter os quantis desejados, bastando ajustar o argumento k. Note que a função retorna as separatrizes expressas em percentis, onde 0% equivale ao mínimo e 100% ao máximo.

```
hosp <- read.table('http://www.filipezabala.com/data/hospital.txt', header = T)
options(digits = 3)                                  # Para melhorar a apresentação
quantile(hosp$altura, probs = seq(0, 1, 1/2))        # Mediana
```

```
##   0%  50% 100%
## 1.51 1.62 1.74
```

```
quantile(hosp$altura, probs = seq(0, 1, 1/3))        # Tercis
```

```
##    0% 33.3% 66.7%  100%
##  1.51  1.61  1.65  1.74
```

```
quantile(hosp$altura, probs = seq(0, 1, 1/4))        # Quartis
```

```
##   0%  25%  50%  75% 100%
## 1.51 1.60 1.62 1.65 1.74
```

```
quantile(hosp$altura, probs = seq(0, 1, 1/10))       # Decis
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
## 1.51 1.57 1.59 1.60 1.62 1.62 1.64 1.65 1.66 1.68 1.74
```

Exercício 2.5. Interprete os quantis do Exemplo 2.31.

Exercício 2.6. Considere as separatrizes discutidas nesta Seção.
a. Verifique que as separatrizes mediana (Md), segundo quartil (Q_2) são equivalentes.
b. Existem outras medidas equivalentes às do item (a)? Justifique.
c. Considere algum k diferente dos apresentados e atribua um nome e uma simbologia.
d. Se existem k 'fatias', quantas são as separatrizes?

Exercício 2.7. Utilizando a função `quantile` calcule as separatrizes discutidas nesta Seção com os dados da coluna `filhos` disponível em http://www.filipezabala.com/data/hospital.txt.

2.4 Medidas de Dispersão (ou Variabilidade)

2.4.1 Amplitude

A *amplitude* é a medida de dispersão mais simples de ser calculada, e fornece uma informação rápida sobre a variabilidade do conjunto de dados. É calculada pela expressão

$$A = \max X - \min X \tag{2.15}$$

Exemplo 2.32. (Amplitude com valores positivos) A amplitude das temperaturas 6, 4, 9, 20, 7 e 12 é

$$A = 20 - 4 = 16.$$

```
temp <- c(6,4,9,20,7,12)  # dados
max(temp)-min(temp)       # pela Eq. (2.15)
```

```
## [1] 16
```

```
A <- range(temp)          # a função 'range' retorna o mínimo e o máximo
diff(A)                   # a função 'diff' calcula a diferença
```

```
## [1] 16
```

Exemplo 2.33. (Amplitude com valores negativos) A amplitude das temperaturas 6, -4, 9, 20, 7 e 12 é

$$A = 20 - (-4) = 24.$$

```
temp <- c(6,-4,9,20,7,12) # dados
diff(range(temp))         # funções aninhadas ('nested functions')
```

```
## [1] 24
```

2.4.2 Variância

A *variância* é a principal medida de dispersão da Estatística. É uma média quadrática em relação à média, i.e., avalia o quanto, em média, os dados variam ao quadrado em torno da média. A variância *universal* pode ser calculada pelas Equações (2.16) e (2.17).

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{2.16}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2 \tag{2.17}$$

Exemplo 2.34. A variância universal do conjunto de dados 186, 402, 191, 20, 7 e 124 é

Equação (2.16)

$$\sigma^2 = \frac{\sum_{i=1}^{6}(x_i - 155)^2}{6} = \frac{(186-155)^2 + (402-155)^2 + \cdots + (124-155)^2}{6} = \frac{104356}{6} = 17392.\bar{6}$$

Equação (2.17)

$$\sigma^2 = \frac{186^2 + 402^2 + 191^2 + 20^2 + 7^2 + 124^2}{6} - 155^2 = \frac{248506}{6} - 24025 = 17392.\bar{6}$$

```
(var.p <- var(c(186,402,191,20,7,124))*(5/6))    # variância amostral*(1/fator de correção)
```

```
## [1] 17393
```

A variância *amostral* pode ser calculada pelas Equações (2.18) e (2.19)

$$\hat{\sigma}^2 = s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2.18}$$

$$\hat{\sigma}^2 = s_n^2 = \left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2\right)\left(\frac{n}{n-1}\right) \tag{2.19}$$

Exemplo 2.35. A variância amostral do conjunto de dados 186, 402, 191, 20, 7 e 124 é

Equação (2.18)

$$s_6^2 = \frac{\sum_{i=1}^{6}(x_i - 155)^2}{6-1} = \frac{(186-155)^2 + (402-155)^2 + \cdots + (124-155)^2}{6-1} = \frac{104356}{5} = 20871.2$$

Equação (2.19)

$$s_6^2 = \left(\frac{186^2 + 402^2 + 191^2 + 20^2 + 7^2 + 124^2}{6} - 155^2\right)\left(\frac{6}{5}\right) = 17392.\bar{6}\times 1.2 = 20871.2$$

```
(var.a <- var(c(186,402,191,20,7,124)))     # 'var' calcula a variância amostral
```

```
## [1] 20871
```

Assim, se o conjunto de dados deste exemplo representar uma amostra observada em 6 vezes que se contou o número de passos até a lixeira mais próxima na capital do Rio Grande (do Sul), pode-se dizer que a variância amostral é 20871.2 passos^2. Dica: não tente interpretar este valor.

Note pela Equação (2.18) que a variância amostral é dividida por $n-1$ e não por n. Isto faz com que a variância amostral seja maior ou igual à variância universal para os mesmos dados. Intuitivamente pode-se pensar como uma espécie de penalidade aplicada a esta medida quando observa-se apenas parte do universo (amostra). Da mesma forma pode-se pensar na variância amostral como o produto entre a variância universal σ^2 e o fator $n/(n-1)$, descrito por

$$s_n^2 = \sigma^2\left(\frac{n}{n-1}\right) \tag{2.20}$$

2.4.3 Desvio Padrão

O desvio padrão é a raiz quadrada da variância. O motivo de calcular o desvio padrão é que a sua interpretação é mais intuitiva se comparada à da variância, uma vez que a unidade de medida do desvio padrão é a mesma da variável X. As fórmulas do desvio padrão universal e amostral são dadas respectivamente pelas equações[8] (2.21) e (2.22).

$$\sigma = \sqrt{\sigma^2} \tag{2.21}$$

$$s_n = \sqrt{s_n^2} \tag{2.22}$$

[8] Se você ficou confuso com a notação, escreva $\sigma^2 = V$ e $\sigma = D$ (bem como $s^2 = v$ e $s = d$) e repense o problema.

Exemplo 2.36. (Desvio padrão universal) Do Exemplo 2.34 sabe-se que a variância universal do conjunto de dados 186, 402, 191, 20, 7 e 124 é $\sigma^2 = 17392.\bar{6}$. Assim, o desvio padrão universal é

$$\sigma = \sqrt{17392.\bar{6}} \approx 131.88126.$$

```
dat <- c(186,402,191,20,7,124)     # dados
(dp.p <- sd(dat) * sqrt(5/6))      # s_n * raiz(1/fator de correção)
```

```
## [1] 132
```

```
all.equal(dp.p, sqrt(var.p))       # 'dp.p' é igual à raiz quadrada de 'var.p'
```

```
## [1] TRUE
```

```
all.equal(dp.p^2, var.p)           # 'dp.p' ao quadrado é igual a 'var.p'
```

```
## [1] TRUE
```

Exemplo 2.37. Do Exemplo 2.35 sabe-se que a variância amostral do conjunto de dados 186, 402, 191, 20, 7 e 124 é $s_6^2 = 20871.2$. Assim, o desvio padrão amostral é

$$s_6 = \sqrt{20871.2} \approx 144.46868.$$

```
dat <- c(186,402,191,20,7,124)     # dados
(dp.a <- sd(dat))                  # 'sd' calcula o desvio padrão amostral
```

```
## [1] 144
```

```
all.equal(dp.a, sqrt(var.a))       # 'dp.a' é igual à raiz quadrada de 'var.a'
```

```
## [1] TRUE
```

```
all.equal(dp.a^2, var.a)           # 'dp.a' ao quadrado é igual a 'var.a'
```

```
## [1] TRUE
```

Assim, se o conjunto de dados deste exemplo representar uma amostra observada em 6 vezes que se contou o número de passos até a lixeira mais próxima na capital do Rio Grande (do Sul), pode-se dizer que o desvio padrão (amostral, claro) é de aproximadamente 144.5 passos. Pode-se pensar neste valor como uma oscilação média aproximada em torno da média aritmética.

2.4.4 Coeficiente de variação

O *coeficiente de variação* é uma medida de comparação de variabilidades, uma vez que ajusta o desvio padrão pela média. É preferível ao desvio padrão por ser um número adimensional, i.e., não possui unidade de medida, tornando quaisquer conjuntos de dados comparáveis em termos de variabilidade. É utilizado

em diversas áreas da Estatística, mas é popularmente conhecido como medida de risco em carteiras de ativos.

As fórmulas do coeficiente de variação universal e amostral são dadas respectivamente pelas equações (2.23) e (2.24).

$$\gamma = \frac{\sigma}{\mu} \tag{2.23}$$

$$\hat{\gamma} = g = \frac{s}{\bar{x}} \tag{2.24}$$

Exemplo 2.38. (Coeficiente de variação) Duas variáveis são obtidas em um certo experimento químico. A variável X é medida em microgramas e possui média de 0.0045 μg e desvio padrão de 0.0056 μg. A variável Y é medida em mols e possui média de 3549 mols e desvio padrão de 419 mols. O coeficiente de variação de X é dado por $g_X = \frac{0.0056}{0.0045} \approx 1.24$, e de Y por $g_Y = \frac{419}{3549} \approx 0.12$. Portanto, como $1.24 > 0.12$, conclui-se que o conjunto de dados X varia mais do que Y.

```
mx <- 0.0045
dx <- 0.0056
round(gx <- dx/mx, 2)   # Coeficiente de variação de X
```

```
## [1] 1.24
```

```
my <- 3549
dy <- 419
round(gy <- dy/my, 2)   # Coeficiente de variação de Y
```

```
## [1] 0.12
```

2.5 Outras medidas

2.5.1 Assimetria (ou Obliquidade)

Assimetria ou *obliquidade* é uma medida que avalia a assimetria de uma distribuição de frequência. Existem diversas definições na literatura, das quais apresentam-se três alternativas.

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]^{3/2}} \tag{2.25}$$

$$b_1 = g_1\left(\frac{n-1}{n}\right)^{3/2} = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]^{3/2}} \tag{2.26}$$

$$G_1 = g_1\sqrt{\frac{n(n-1)}{n-2}} = b_1\frac{n^2}{(n-1)(n-2)} \tag{2.27}$$

```
set.seed(1); x <- rnorm(100)     # Gerando 100 valores N(0,1) com semente fixa
e1071::skewness(x, type = 1)     # Definição clássica de assimetria, Equação (2.25)
```

```
## [1] -0.0722
```

```
e1071::skewness(x, type = 2)     # Utilizada no SAS, SPSS e Excel, Equação (2.27)
```

```
## [1] -0.0733
```

```
e1071::skewness(x, type = 3)     # Padrão do R, utilizada no MINITAB e BMDP, Equação (2.26)
```

```
## [1] -0.0712
```

2.5.2 Curtose

A *curtose* é uma medida de achatamento de uma distribuição de frequência. Assim como na assimetria, das diversas definições de curtose apresentam-se três alternativas.

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]^2} - 3 \tag{2.28}$$

$$b_2 = (g_2 + 3)\left(1 - \frac{1}{n}\right)^2 - 3 = \frac{m_4}{s^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^4}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]^2} - 3 \tag{2.29}$$

$$G_2 = \frac{[(n+1)g_2 + 6]\,(n-1)}{(n-2)(n-3)} \tag{2.30}$$

```
set.seed(1); x <- rnorm(100)     # Gerando 100 valores N(0,1) com semente fixa
e1071::kurtosis(x, type = 1)     # Definição clássica de curtose, Equação (2.28)
```

```
## [1] 0.00765
```

```
e1071::kurtosis(x, type = 2)     # Usada no SAS, SPSS e Excel, Equação (2.30)
```

```
## [1] 0.0705
```

```
e1071::kurtosis(x, type = 3)     # Padrão do R, usada também no MINITAB e BMDP, Eq. (2.29)
```

```
## [1] -0.0522
```

2.6 Visualização

Visualização é o processo de representar informações ou ideias através de diagramas, gráficos e outros métodos de apresentação visual. De um modo geral, as ferramentas de visualização devem ser claras para o leitor, devendo-se evitar detalhes desnecessários. Um bom visualizador transmite a informação desejada de forma clara, precisa e eficiente. Ao apresentar informação de maneira visual deve-se considerar que 'o menos é mais'. Edward Tufte, "o Leonardo da Vinci dos dados" segundo *The New York Times*, ou "o Galileu dos gráficos" de acordo com a *Bloomberg*, possui uma vasta obra sobre o tema. Em (Tufte and Graves-Morris, 1983), (Tufte, 1993), (Tufte et al., 1998) e (Tufte, 2006) apresenta uma série de considerações e exemplos muito ricos e elegantes, algumas indicadas a seguir.

- o olho humano não diferencia muitas tonalidades de cor, por isso é interessante trabalhar com escalas em degradê, diferentes texturas e espessuras de linha
- para apresentar dados categóricos é interessante ordenar as categorias de forma intuitiva para melhor apresentação
- deve-se evitar o agrupamento de dados contínuos em categorias
- é importante manter a notação coerente com o texto

2.6.1 Menu de opções

- https://plot.ly/r/
- https://www.r-graph-gallery.com/
- https://github.com/d3/d3/wiki/Gallery
- http://kateto.net/network-visualization
- https://www.shinyapps.org/apps/RGraphCompendium/index.php
- http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html
- https://d3js.org/ (JavaScript)

Chapter 3

Probabilidade

Muito há para se falar sobre probabilidade desde a troca de correspondências entre Pascal e Fermat em 1654. Segundo Pierre-Simon Laplace[1], 'a teoria das probabilidades é, basicamente, o senso comum reduzido ao cálculo.' Para o matemático italiano Bruno de Finetti[2], 'PROBABILIDADE NÃO EXISTE'. Neste material serão utilizadas as noções axiomática, subjetiva e frequentista de probabilidade, descritas em detalhes na Seção 2.2 de (Press, 2003).

3.1 Propriedades

3.1.1 Propriedades fundamentais (Axiomas de Kolmogorov)

Um leitor mais atento pode perceber que foi feita uma combinação entre os axiomas de Kolmolgorov e as propriedades que são consequências destes axiomas. Tal abordagem tem por finalidade simplificar o entendimento neste curso de nível introdutório. Para maiores detalhes, recomenda-se (James, 2010).

- **P1**

$$0 \leq Pr(A) \leq 1 \tag{3.1}$$

- **P2**

$$Pr(\Omega) = 1 \tag{3.2}$$

- **P3** Se A_1, A_2, ..., A_k são conjuntos disjuntos, então

$$Pr(A_1 \cup A_2 \cup ... \cup A_k) = Pr(A_1) + Pr(A_2) + ... + Pr(A_k) \tag{3.3}$$

[1] "*[L]a théorie des probabilités n'est au fond, que le bon sens réduit au calcul.*", (Laplace, 1825) página 275.

[2] "*PROBABILITY DOES NOT EXIST.*" (de Finetti, 1974) página x.

3.1.2 Propriedades secundárias

Das propriedades fundamentais resultam outras, apresentadas sem demonstração:

- **P4**

$$Pr(A) = 1 - Pr(A^c) \tag{3.4}$$

- **P5**

$$Pr(\emptyset) = 0 \tag{3.5}$$

- **P6**
  Se A_1 e A_2 são dois conjuntos quaisquer, então

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B) \tag{3.6}$$

- **P7**

$$Pr([A \cup B]^c) = Pr(A^c \cap B^c) \tag{3.7}$$

- **P8**

$$Pr([A \cap B]^c) = Pr(A^c \cup B^c) \tag{3.8}$$

3.2 R como um conjunto de tabelas estatísticas

(Venables et al., 2020) apontam que um uso conveniente de R é fornecer um conjunto abrangente de tabelas estatísticas. Funções são fornecidas para avaliar a função densidade de probabilidade (FDP) $f(x)$, a função distribuição acumulada (FDA) $F(x) = Pr(X \leq x)$, a função quantil (dado q, o menor x tal que $Pr(X \leq x) > q$) e também para simular valores das distribuições. Utiliza-se o prefixo `d` para a densidade, `p` para o FDA, `q` para a função de quantil e `r` para simulação pseudo-aleatória. A seguir são apresentadas as distribuições de probabilidade disponíveis no *base R*. Para mais distribuições podem-se utilizar os pacotes adicionais `mvtnorm` (normal e t multivariadas) e `VGAM` (Dirichlet, multinomial, beta-binomial, entre outras).

Distribution	R name	additional arguments
beta	`beta`	`shape1, shape2, ncp`
binomial	`binom`	`size, prob`
Cauchy	`cauchy`	`location, scale`
chi-squared	`chisq`	`df, ncp`
exponential	`exp`	`rate`
F	`f`	`df1, df2, ncp`
gamma	`gamma`	`shape, scale`
geometric	`geom`	`prob`
hypergeometric	`hyper`	`m, n, k`
log-normal	`lnorm`	`meanlog, sdlog`
logistic	`logis`	`location, scale`
negative binomial	`nbinom`	`size, prob`
normal	`norm`	`mean, sd`
Poisson	`pois`	`lambda`
signed rank	`signrank`	`n`
Student's t	`t`	`df, ncp`
uniform	`unif`	`min, max`
Weibull	`weibull`	`shape, scale`
Wilcoxon	`wilcox`	`m, n`

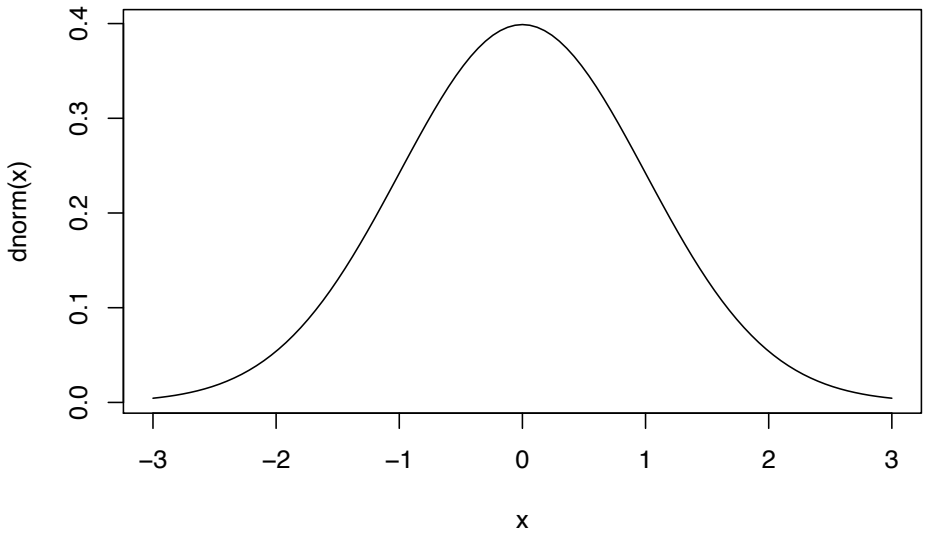
3.3 Distribuição Normal

3.3.1 Normal univariada

A distribuição normal univariada é dada pela expressão

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$

```
# gráfico da densidade da normal padrão, N(0,1)
curve(dnorm(x), xlim = c(-3,3))
```




```
# distribuição acumulada
pnorm(0)
```

```
## [1] 0.5
```

```
pnorm(1.645)
```

```
## [1] 0.95
```

```
pnorm(1.96)
```

```
## [1] 0.975
```

```
# quantis (separatrizes)
qnorm(0.5)
```

```
## [1] 0
```

```
qnorm(0.95)
```

```
## [1] 1.64
```

```
qnorm(0.975)
```

```
## [1] 1.96
```

3.3.2 Normal bivariada

No caso bivariado pode-se definir

$$f(x_1, x_2|\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}.$$

Exercício 3.1. Verifique que o produto de duas normais univariadas equivale à definição no caso bivariado quando $\rho = 0$.

```
library(rgl)

# parâmetros
n <- 100
x1 <- seq(-5, 5, length = n)
x2 <- seq(-5, 5, length = n)
m1 <- 0
m2 <- 0
s1 <- 1  # desvio padrão
s2 <- 2

# produto de normais independentes, \rho = 0
z1 <- outer(x1, x2, function(x,y) dnorm(x,m1,s1) * dnorm(y,m2,s2))

# gráficos
persp3d(x1, x2, z1, col = 'gray')
rglwidget()
```

```r
library(rgl)
library(mvtnorm)

# parâmetros
n <- 100
x1 <- seq(-5, 5, length = n)
x2 <- seq(-5, 5, length = n)
m1 <- 0
m2 <- 0
s1 <- 1^2  # variância
s2 <- 2^2

# via mvtnorm::dmvnorm, \rho = 0
m <- c(m1,m2)
s <- diag(c(s1,s2))
z2 <- outer(x1, x2, function(x,y) dmvnorm(cbind(x,y), mean = m, sigma = s))

# gráficos
persp3d(x1, x2, z2, col = 'lightblue')
rglwidget()
```

```r
library(rgl)
library(mvtnorm)

# parâmetros
n <- 100
x1 <- seq(-5, 5, length = n)
x2 <- seq(-5, 5, length = n)
m1 <- 0
m2 <- 0
s1 <- 1^2
s2 <- 2^2
r <- 0.9

# via mvtnorm::dmvnorm, \rho = 0.9
m <- c(m1,m2)
s <- matrix(c(s1,r,r,s2), nrow = 2, byrow = T)
z3 <- outer(x1, x2, function(x,y) dmvnorm(cbind(x,y), mean = m, sigma = s))

# gráficos
persp3d(x1, x2, z3, col = 'lightgreen')
rglwidget()
```

As probabilidades podem ser calculadas através da função `pmvnorm` do pacote `mvtnorm`.

```
library(mvtnorm)

# parâmetros
m <- c(0,0)
s <- diag(2)

# Pr(X1 < 0, X2 < 0)
lower <- c(-Inf, -Inf)
upper <- c(0, 0)
pmvnorm(lower, upper, m, s)
```

```
## [1] 0.25
## attr(,"error")
## [1] 1e-15
## attr(,"msg")
## [1] "Normal Completion"
```

Exercício 3.2. Interprete o valor 0.25 calculado no exemplo acima.

3.3.3 Normal multivariada

Para o caso multivariado define-se

$$f(x|\mu,\Sigma) = \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} \exp\left\{ -\tfrac{1}{2}(x-\mu)'\Sigma^{1/2}(x-\mu)\right\},$$

$$\mu = (\mu_1, \mu_2, \ldots, \mu_p)', \; \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}.$$

Chapter 4

Amostragem

Definição 4.1. *Amostragem* é o processo de obtenção de uma amostra.

Inicia com o *plano amostral*, uma avaliação que leva em conta as medidas a serem avaliadas e os recursos disponíveis. Da mesma forma que os laboratórios retiram amostras de sangue para exames de saúde, cozinheiros experimentam parte da comida para provar os temperos e fábricas realizam testes destrutivos em parte da sua produção para avaliar a qualidade do que está sendo produzido. Será feita uma breve revisão dos principais conceitos de amostragem baseada em (Bolfarine and Bussab, 2005).

4.1 Definições básicas

4.1.1 Unidade Elementar

Definição 4.2. A *unidade elementar*, *unidade populacional* ou simplesmente *elemento* é a entidade portadora das informações que pretende-se coletar.

A unidade elementar pode ser um objeto, animal ou pessoa. Em certos casos existe mais de uma maneira de definir a unidade elementar, onde se faz necessário o entendimento dos especialistas envolvidos. A unidade elementar é uma das definições mais importantes do campo científico, pois é base de toda a construção das hipóteses de pesquisa.

Exemplo 4.1. (Pesquisa eleitoral I) Em uma pesquisa eleitoral, classifica-se o eleitor como unidade elementar.

4.1.2 Unidade Amostral

Definição 4.3. A *unidade amostral* é uma composição de uma ou mais unidades elementares.

Exemplo 4.2. (Pesquisa eleitoral II) Em uma pesquisa eleitoral na rua, o eleitor é também unidade amostral. Caso as entrevistas sejam feitas de casa em casa, o domicílio passa a ser unidade amostral, i.e., um conjunto de unidades elementares (eleitores).

4.1.3 Sistema de referências

Em relação às informações de um estudo, deve-se fazer inicialmente uma avaliação das bases de dados já disponíveis para então proceder com a avaliação da viabilidade de um levantamento de dados mais específico. Este levantamento envolve se montar um plano amostral, contratar, treinar e manter pessoas para a coleta, criar os protocolos de resposta bem como gerar e analisar os bancos de dados. Caso decida-se realizar tal levantamento, serão necessárias listas relacionando as unidades populacionais e amostrais. Na falta de tais listas, utilizam-se *sistemas de referências*, que são fontes que descrevem o universo a ser investigado. Podem ser informações razoavelmente atualizadas, como mapas, censos ou listas reunidas.

4.2 Universo $\mathcal{U}$

Definição 4.4. *Universo* ou *população* é o conjunto de todas as unidades elementares de interesse.

Usualmente o universo possui tamanho N elevado, até mesmo infinito, mas em alguns casos pode ser relativamente pequeno. É denotado formalmente por

$$\mathcal{U} = \{1, 2 \dots, N\}.$$

Exemplo 4.3. (Pesquisa eleitoral III) Em 2018 o universo de eleitores do município de Porto Alegre compreendia 1,100,163 eleitores[1], i.e., $N = 1\,100\,163$. Formalmente

$$\mathcal{U} = \{1, 2 \dots, 1\,100\,163\}.$$

Definição 4.5. *Elemento universal*, *elemento populacional* ou *unidade elementar* denota um elemento $i \in \mathcal{U}$.

[1] Fonte: `http://www.tse.jus.br/eleicoes/estatisticas/estatisticas-eleitorais`.

Definição 4.6. *Característica(s) de interesse* denota(m) a variável ou o conjunto de k variáveis associada(o) a cada elemento do universo, anotado por $X = (x_1, x_2, ..., x_N) = \left(\begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1k} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2k} \end{bmatrix}, ..., \begin{bmatrix} x_{N1} \\ x_{N2} \\ \vdots \\ x_{Nk} \end{bmatrix} \right) = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{Nk} \end{pmatrix}$.

Exemplo 4.4. Considere que no universo $\mathcal{U} = \{1, 2, 3\}$ de tamanho $N = 3$ o sujeito 1 seja do sexo feminino com 24 anos de idade e 1.66m de altura, o sujeito 2 do sexo masculino com idade de 32 anos e 1.81m de altura, e o sujeito 3 do sexo masculino com 49 anos com altura de 1.73m. Assim,

$$X = (x_1, x_2, x_3) = \left(\begin{bmatrix} 24 \\ 1.66 \\ F \end{bmatrix}, \begin{bmatrix} 32 \\ 1.81 \\ M \end{bmatrix}, \begin{bmatrix} 49 \\ 1.73 \\ M \end{bmatrix} \right) = \begin{pmatrix} 24 & 32 & 49 \\ 1.66 & 1.81 & 1.73 \\ M & F & M \end{pmatrix}.$$

4.2.1 Parâmetros

Definição 4.7. *Parâmetro universal* ou *parâmetro populacional* denota uma função ou medida que depende de todas as características de interesse.

Exemplo 4.5. O parâmetro *total universal* é dado por

$$\tau = \sum_{i=1}^{N} x_i. \tag{4.1}$$

Exemplo 4.6. O parâmetro *média universal* é dado por

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{\tau}{N}. \tag{4.2}$$

Exemplo 4.7. Uma variável é chamada *dicotômica* quando assume apenas dois possíveis valores tais como sim/não, verdadeiro/falso, ligado/desligado, etc. A característica de interesse é chamada *sucesso* e a outra característica de *fracasso*. Por conveniência associa-se o sucesso ao valor $x = 1$ e fracasso a $x = 0$. Desta forma simboliza-se $\sum_{i=1}^{N} x_i$ como o total de sucessos observados no universo. Nesta situação o parâmetro *proporção universal* é dado por

$$\pi = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{4.3}$$

Exemplo 4.8. O parâmetro *variância universal* é dado pelas Equações (2.16) e (2.17).

Exemplo 4.9. O parâmetro *desvio padrão universal* é a raiz quadrada da variância universal, dado pela Equação (2.21).

Exemplo 4.10. O parâmetro *covariância universal* é dado por

$$\sigma_{XY} = Cov[X, Y] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y). \tag{4.4}$$

Exemplo 4.11. O parâmetro *correlaçao universal* é dado por

$$\rho_{XY} = Cor[X, Y] = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{4.5}$$

Exercício 4.1. Utilizando os dados do Exemplo 4.4, calcule os parâmetros dos Exemplos 4.5 a 4.11.

Exercício 4.2. Mostre que as Equações (2.16) e (2.17) são equivalentes.

4.3 Amostras

Definição 4.8. Considere o universo $\mathcal{U} = \{1, 2, \ldots, N\}$. Uma *amostra* é qualquer sequência de n unidades de $\mathcal{U}$, formalmente denotada por

$$a = (a_1, \ldots, a_n),$$

onde o i-ésimo compomente de a é tal que $a_i \in \mathcal{U}$.

Exemplo 4.12. Seja $\mathcal{U} = \{1, 2, 3\}$. Os vetores $a_A = (2, 3)$, $a_B = (3, 3, 1)$, $a_C = (2)$, $a_D = (2, 2, 3, 3, 1)$ são possíveis amostras de $\mathcal{U}$.

Exemplo 4.13. No Exemplo 4.12, note os tamanhos de amostra $n_A = n(a_A) = 2$, $n_B = n(a_B) = 3$, $n_C = n(a_C) = 1$ e $n_D = n(a_D) = 5$.

Definição 4.9. Seja $\mathcal{A}(\mathcal{U})$ ou simplesmente $\mathcal{A}$ o conjunto de todas as amostras de $\mathcal{U}$, de qualquer tamanho, e $\mathcal{A}_n(\mathcal{U})$ ou simplesmente $\mathcal{A}_n$ a subclasse das as amostras de tamanho n.

Exemplo 4.14. Se $\mathcal{U} = \{1, 2, 3\}$,

$$\mathcal{A}(\mathcal{U}) = \{(1), (2), (3), (1, 1), (1, 2), (1, 3), (2, 1), \dots, (3, 1, 2, 2, 1), \dots\},$$

$$\mathcal{A}_1(\mathcal{U}) = \{(1), (2), (3)\},$$

$$\mathcal{A}_2(\mathcal{U}) = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}.$$

Simplificadamente

$$\mathcal{A} = \{1, 2, 3, 11, 12, 13, 21, \dots, 31221, \dots\},$$

$$\mathcal{A}_1 = \{1, 2, 3\},$$

$$\mathcal{A}_2 = \{11, 12, 13, 21, 22, 23, 31, 32, 33\}.$$

Exemplo 4.15. No exemplo anterior, note o número de elementos (cardinalidade) de cada conjunto:

$$|\mathcal{U}| = 3$$

$$|\mathcal{A}(\mathcal{U})| = \infty$$

$$|\mathcal{A}_1(\mathcal{U})| = 3^1 = 3$$

$$|\mathcal{A}_2(\mathcal{U})| = 3^2 = 9$$

$$\vdots$$

$$|\mathcal{A}_n(\mathcal{U})| = |\mathcal{U}|^n.$$

4.3.1 Plano Amostral

Definição 4.10. Um *plano amostral (ordenado)* é uma função $P(a)$ definida em $\mathcal{A}(\mathcal{U})$ satisfazendo

$$P(a) \geq 0, \ \forall a \in \mathcal{A}(\mathcal{U}),$$

tal que

$$\sum_{a \in \mathcal{A}} P(a) = 1.$$

Exemplo 4.16. Considere $\mathcal{U} = \{1, 2, 3\}$ e $\mathcal{A}(\mathcal{U})$ conforme Exemplo 4.14. É possivel criar infinitos planos amostrais, tais como:

- Plano A · Amostragem Aleatória Simples *com* reposição (AASc)

$P(11) = P(12) = P(13) = 1/9 P(21) = P(22) = P(23) = 1/9 P(31) = P(32) = P(33) = 1/9 P(a) = 0, \ \forall a \in \mathcal{A}(\mathcal{U}).$

- **Plano B · Amostragem Aleatória Simples *sem* reposição (AASs)**

  $P(12) = P(13) = 1/6 P(21) = P(23) = 1/6 P(31) = P(32) = 1/6 P(a) = 0, \ \forall a \in \mathcal{A}(\mathcal{U}).$

- **Plano C · Combinações**

$$P(12) = P(13) = P(23) = 1/3P(a) = 0,\ \forall a \in \mathcal{A}(\mathcal{U}).$$

- **Plano D**

$$P(3) = 9/27P(12) = P(23) = 3/27P(111) = P(112) = P(113) = P(123) = 1/27P(221) = P(222) =$$

Exemplo 4.17. Considere a amostra $a = (1, 2)$ obtida do universo descrito Exemplo 4.4 a partir de algum plano amostral válido. Se o sujeito 1 tem 24 anos de idade e 1.66m de altura, e o sujeito 2 tem 32 anos de idade altura de 1.81m,

$$x = (x_1, x_2) = \left(\begin{bmatrix} 24 \\ 1.66 \end{bmatrix}, \begin{bmatrix} 32 \\ 1.81 \end{bmatrix} \right) = \begin{pmatrix} 24 & 32 \\ 1.66 & 1.81 \end{pmatrix}.$$

Definição 4.11. Uma *estatística* é uma função dos dados amostra a anotada por $h(x)$, i.e., qualquer medida numérica calculada a partir dos valores observados na amostra.

Exemplo 4.18. Considere x, a matriz dos dados da amostra $a = (1, 2)$. São exemplos de estatísticas:

$$h_1 = \frac{24 + 32}{2} = 28 \quad \text{(média das idades)}$$

$$h_2 = \frac{1.66 + 1.81}{2} = 1.735 \quad \text{(média das alturas)}$$

$$h_3 = 32 - 24 = 8 \quad \text{(amplitude das idades)}$$

$$h_4 = \sqrt{24^2 + 32^2} = \sqrt{1600} = 40 \quad \text{(norma das idades)}$$

Exercício 4.3. Calcule as estatísticas do Exemplo 4.18 considerando as amostras $a = (1, 3)$ e $a = (2, 3)$.

4.3.2 Distribuições amostrais

Definição 4.12. A *distribuição amostral* de uma estatística $h(x)$ segundo um plano amostral λ, é a distribuição de probabilidades $H(x)$ definida sobre $\mathcal{A}_\lambda$, com função de probabilidade

$$p_h = P_\lambda(H(x) = h) = P(h) = \frac{f_h}{|\mathcal{A}_\lambda|}.$$

Exemplo 4.19. Considere a variável `idade` do Exemplo 4.4 e as estatísticas $h_1(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $h_2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - h_1(x))^2$ aplicadas sobre o plano amostral A do Exemplo 4.16. Note que $h_1(x)$ e $h_2(x)$ são respectivamente a média e a variância amostrais.

- **Plano A · Amostragem Aleatória Simples *com* reposição (AASc)**

i	1	2	3	4	5	6	7	8	9
a	11	12	13	21	22	23	31	32	33
$P(a)$	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
x	(24,24)	(24,32)	(24,49)	(32,24)	(32,32)	(32,49)	(49,24)	(49,32)	(49,49)
$h_1(x)$	24.0	28.0	36.5	28.0	32.0	40.5	36.5	40.5	49.0
$h_2(x)$	0.0	32.0	312.5	32.0	0.0	144.5	312.5	144.5	0.0

h_1	24.0	28.0	32.0	36.5	40.5	49.0	**Total**
f_{h1}	1	2	1	2	2	1	**9**
p_{h1}	1/9	2/9	1/9	2/9	2/9	1/9	**1**

h_2	0.0	32.0	144.5	312.5	**Total**
f_{h2}	3	2	2	2	**9**
p_{h2}	3/9	2/9	2/9	2/9	**1**

Exemplo 4.20. Considere novamente a variável `idade` do Exemplo 4.4 e a estatística $h_1(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$, agora aplicada sobre o plano amostral B do Exemplo 4.16.

- **Plano B · Amostragem Aleatória Simples *sem* reposição (AASs)**

i	1	2	3	4	5	6
a	12	13	21	23	31	32
$P(a)$	1/6	1/6	1/6	1/6	1/6	1/6
x	(24,32)	(24,49)	(32,24)	(32,49)	(49,24)	(49,32)
$h_1(x)$	28.0	36.5	28.0	40.5	36.5	40.5

h_1	28.0	36.5	40.5	**Total**
f_{h1}	2	2	2	**6**

p_{h1}	2/6	2/6	2/6	**1**

Exemplo 4.21. Considere novamente a variável `idade` do Exemplo 4.4 e a estatística $h_1(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$, agora aplicada sobre o plano amostral C do Exemplo 4.16.

- Plano C · Combinações

i	1	2	3
a	12	13	23
$P(a)$	1/3	1/3	1/3
x	(24,32)	(24,49)	(32,49)
$h_1(x)$	28.0	36.5	40.5

h_1	28.0	36.5	40.5	**Total**
f_{h1}	1	1	1	**3**
p_{h1}	1/3	1/3	1/3	**1**

Exercício 4.4. Refaça os Exemplos 4.19, 4.20 e 4.21 considerando a variável `altura`. Para os Exemplos 4.20 e 4.21, calcule também a estatística $h_2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - h_1(x))^2$.

Exemplo 4.22. A seguir são implementadas em R as resoluções dos Exemplos 4.19 e 4.20.

```
U <- 1:3                      # universo
(aasc <- expand.grid(U,U))    # AASc de tamanho n=2
```

```
##   Var1 Var2
## 1    1    1
## 2    2    1
## 3    3    1
## 4    1    2
```

```
## 5    2    2
## 6    3    2
## 7    1    3
## 8    2    3
## 9    3    3
```

```
(aasc <- cbind(aasc[,2],aasc[,1])) # trocando as colunas para melhor leitura
```

```
##       [,1] [,2]
##  [1,]    1    1
##  [2,]    1    2
##  [3,]    1    3
##  [4,]    2    1
##  [5,]    2    2
##  [6,]    2    3
##  [7,]    3    1
##  [8,]    3    2
##  [9,]    3    3
```

```
(aass <- aasc[-c(1,5,9),]) # AASs de tamanho n=2
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    2    1
## [4,]    2    3
## [5,]    3    1
## [6,]    3    2
```

```
x1 <- c(24,32,49)           # dados de idade
n <- ncol(aasc)
# AASc
(xc <- cbind(x1[aasc[,1]], x1[aasc[,2]])) # dados amostrais de idade com reposição
```

```
##       [,1] [,2]
##  [1,]   24   24
##  [2,]   24   32
##  [3,]   24   49
##  [4,]   32   24
##  [5,]   32   32
##  [6,]   32   49
##  [7,]   49   24
##  [8,]   49   32
##  [9,]   49   49
```

```
(mxc <- rowMeans(xc))       # estatística h1(x) aplicada na AASc
```

```
## [1] 24.0 28.0 36.5 28.0 32.0 40.5 36.5 40.5 49.0
```

```
(tabc <- table(mxc))        # frequência amostral de h1(y) aplicada na AASc
```

```
## mxc
##   24   28   32 36.5 40.5   49
##    1    2    1    2    2    1
```

```
MASS::fractions(prop.table(tabc)) # distribuição amostral de h1(x) aplicada na AASc
```

```
## mxc
##   24   28   32 36.5 40.5   49
##  1/9  2/9  1/9  2/9  2/9  1/9
```

```
# vyc <- (rowMeans(xc^2)-mxc^2)*(n/(n-1))
# AASs
(xs <- cbind(x1[aass[,1]], x1[aass[,2]])) # dados amostrais de idade sem reposição
```

```
##      [,1] [,2]
## [1,]   24   32
## [2,]   24   49
## [3,]   32   24
## [4,]   32   49
## [5,]   49   24
## [6,]   49   32
```

```
(mxs <- rowMeans(xs))       # estatística h(x) aplicada na AASs
```

```
## [1] 28.0 36.5 28.0 40.5 36.5 40.5
```

```
(tabs <- table(mxs))        # frequência amostral de h(x) aplicada na AASs
```

```
## mxs
##   28 36.5 40.5
##    2    2    2
```

```
MASS::fractions(prop.table(tabs)) # distribuição amostral de h(x) aplicada na AASs
```

```
## mxs
##   28 36.5 40.5
##  1/3  1/3  1/3
```

Exemplo 4.23. As resoluções dos Exemplos 4.20 e 4.21 podem ser implementadas no pacote `arrangements` do R. Note que são obtidas as amostras via AASs através da função `permutations` e as amostras por combinação, sem qualquer tipo de repetição, pela função `combinations`.

```
library(arrangements)
x1 <- c(24,32,49)  # dados de idade
# AASs
npermutations(3,2) # número de AASs
```

```
## [1] 6
```

```
(aass <- permutations(3,2)) # gerando as AASs
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    2    1
## [4,]    2    3
## [5,]    3    1
## [6,]    3    2
```

```
(maass <- matrix(x1[t(aass)], ncol=2, byrow = T))
```

```
##      [,1] [,2]
## [1,]   24   32
## [2,]   24   49
## [3,]   32   24
## [4,]   32   49
## [5,]   49   24
## [6,]   49   32
```

```
rowMeans(maass)
```

```
## [1] 28.0 36.5 28.0 40.5 36.5 40.5
```

```
mean(rowMeans(maass)) # plano amostral não viesado
```

```
## [1] 35
```

```
# Combinações
ncombinations(3,2) # número de amostras via combinação
```

```
## [1] 3
```

```
(comb <- combinations(3,2)) # gerando as amostras via combinação
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    2    3
```

```
(mcomb <- matrix(x1[t(comb)], ncol=2, byrow = T))
```

```
##      [,1] [,2]
## [1,]   24   32
## [2,]   24   49
## [3,]   32   49
```

```
rowMeans(mcomb)
```

```
## [1] 28.0 36.5 40.5
```

```
mean(rowMeans(mcomb)) # plano amostral não viesado
```

```
## [1] 35
```

Desafio 4.1. *Generalize os Exemplos 4.22 e 4.21 para qualquer tamanho de amostra, parametrizando as opções com e sem reposição, bem como para combinações. Por fim, adicione um argumento que permita calcular qualquer estatística.*

Teorema Central do Limite

O *Teorema Central do Limite* (TCL) é um dos principais resultados da Probabilidade. Ele mostra que, sob certas condições razoavelmente alcançadas na prática, a soma ou média de uma sequência de variáveis aleatórias independentes e identicamente distribuídas (*iid*)[2] têm distribuição aproximadamente normal. Este resultado permite a resolução aproximada de problemas que envolvam muitos cálculos, usualmente impraticáveis dado o volume de operações necessárias.

Teorema 4.1. *(Teorema Central do Limite de Lindeberg-Lévy) Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias iid com $E(X_i) = \mu$ e $V(X_i) = \sigma^2$. Considerando $S = X_1 + X_2 + \ldots + X_n$, $M = S/n$ e se $n \longrightarrow \infty$, então*

$$Z = \frac{S - n\mu}{\sigma\sqrt{n}} = \frac{M - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1). \tag{4.6}$$

A *correção de continuidade* ocorre quando soma-se 0.5 no numerador da Equação (4.6). (James, 2010) sugere o uso da expressão 'Teorema Central do Limite' no lugar de 'Teorema do Limite Central', pois central é o teorema, não o limite. A origem da expressão é atribuída ao matemático húngaro George Pólya, ao se referir a *der zentrale Grenzwertsatz*, i.e., o 'central' refere-se ao 'teorema do limite'.

Distribuição amostral da proporção

A proporção é uma média no caso de a variável admitir apenas os valores 0 e 1, portanto o TCL se aplica diretamente a este tipo de estrutura.

Exemplo 4.24. (Aproximação da binomial pela normal) Se considerarmos $n = 420$ lançamentos de uma moeda com $p = 0.5$, temos que a v.a. X: *número de caras* é tal que $X \sim \mathcal{B}(420, 0.5)$. A probabilidade de obtermos até 200 caras pode ser aproximada pelo pelo TCL.

$$Pr(X \leq 200) \approx Pr\left(Z < \frac{200 - 420 \times 0.5}{\sqrt{420 \times 0.5 \times 0.5}}\right) = \Phi(-0.9759) \approx 0.164557.$$

[2] Variáveis que apresentam mesma distribuição de probabilidade com os mesmos parâmetros.

Utilizando a correção de continuidade,

$$Pr(X \leq 200) \approx Pr\left(Z < \frac{200 + 0.5 - 420 \times 0.5}{\sqrt{420 \times 0.5 \times 0.5}}\right) = \Phi(-0.9271) \approx 0.176936.$$

Com um computador é possível calcular a probabilidade exata, perceba a proximidade dos resultados.

$$Pr(X \leq 200) = \left[\binom{420}{0} + \binom{420}{1} + \cdots + \binom{420}{200}\right] 0.5^{420} = 0.1769429.$$

```
n <- 420
p <- 0.5
S <- 200
mS <- n*p  # 210
sS <- sqrt(n*p*(1-p))  # 10.24695
# Aproximação da binomial pela normal SEM correção de continuidade
(z <- (S-mS)/sS)
```

```
## [1] -0.976
```

```
pnorm(z)
```

```
## [1] 0.165
```

```
# Aproximação da binomial pela normal COM correção de continuidade
(zc <- (S+0.5-mS)/sS)
```

```
## [1] -0.927
```

```
pnorm(zc)
```

```
## [1] 0.177
```

```
# Probabilidade exata
pbinom(S,n,p)
```

```
## [1] 0.177
```

Distribuição amostral da média

Com base no Teorema Central do Limite sabe-se que a distribuição das médias amostrais de qualquer variável X que satisfaça as condições do teorema converge para a distribuição normal. Considere que X tem uma distribuição $\mathcal{D}$ qualquer, com média μ e desvio padrão σ, simbolizada por

$$X \sim \mathcal{D}(\mu, \sigma).$$

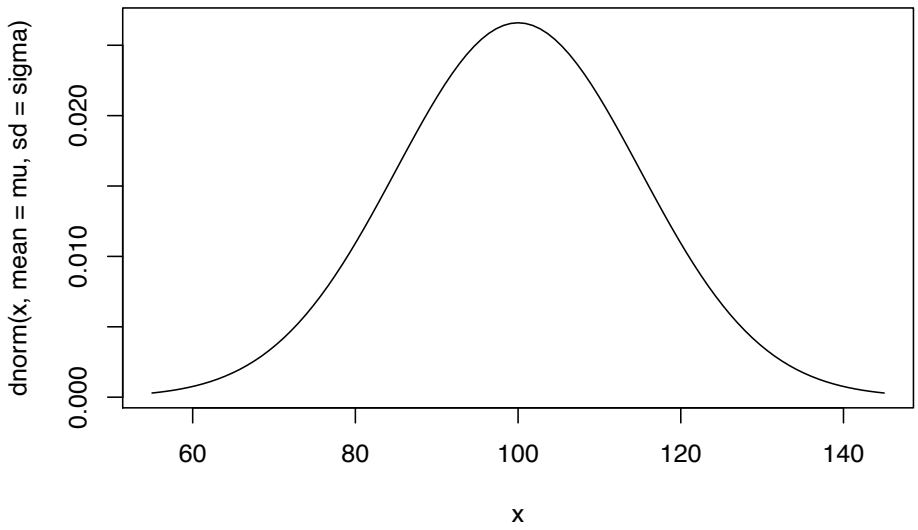
Pelo TCL, a distribuição das médias amostrais de qualqer tamanho n_0 é tal que

$$\bar{X}_{n_0} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n_0}}\right).$$

A medida $\sigma/\sqrt{n_0}$ é conhecida como *erro padrão (da média)*. O TCL é um resultado assintótico[3], portanto quanto mais próxima $\mathcal{D}$ estiver de $\mathcal{N}$, mais rápida deve ser a convergência de $\bar{X}_{n_0}$ para a distribuição normal.

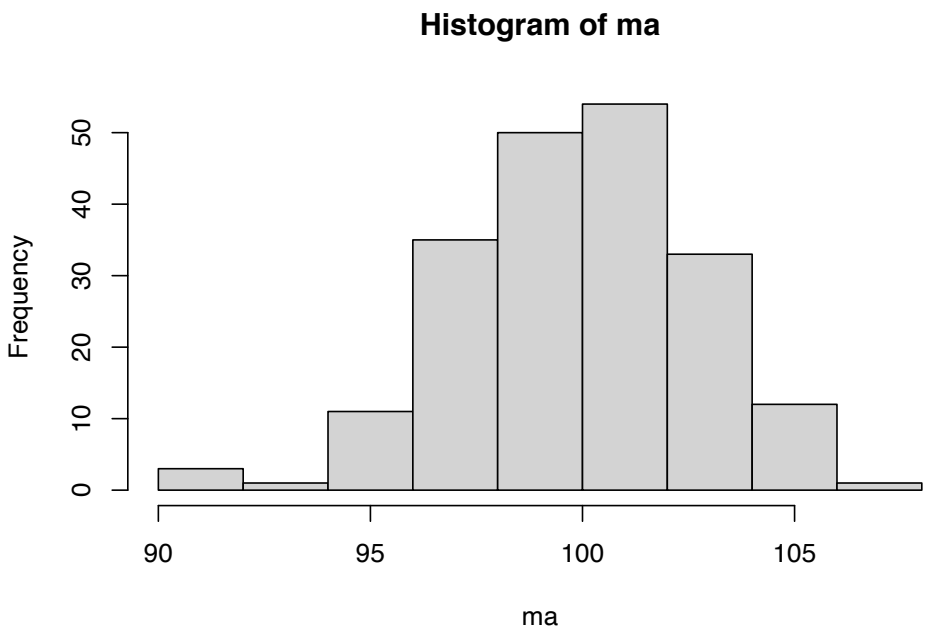
Exemplo 4.25. Considere a variável aleatória X: *QI da população mundial*, admitida com distribuição normal de média $\mu = 100$ e desvio padrão de $\sigma = 15$, anotada por $X \sim \mathcal{N}(100, 15)$.

```
mu <- 100 # média de X
sigma <- 15  # desvio padrão de X
curve(dnorm(x, mean=mu, sd=sigma), from=mu-3*sigma, to=mu+3*sigma) # X ~ N(100,15)
```




```
n0 <- 25 # tamanho das amostras
n <- 200 # número de amostras
set.seed(1234) # fixando semente pseudo-aleatória para garantir replicação
a <- MASS::mvrnorm(n0, mu = rep(mu,n), Sigma = sigma^2*diag(n)) # amostras
ma <- colMeans(a) # médias das n amostras
hist(ma) # histograma das médias
```

[3]Um resultado assintótico é aquele que depende de uma ou mais variávies sendo observadas próximas a certos limites de referência.




```
mean(ma) # média das médias amostrais, próxima de mu
```

```
## [1] 99.9
```

```
sd(ma) # desvio padrão das médias, próximo de sigma/raiz(n0)
```

```
## [1] 2.82
```

```
sigma/sqrt(n0) # sigma/raiz(n0)
```

```
## [1] 3
```

Exercício 4.5. Refaça o Exemplo 4.25 alterando os valores de `n0` e `n`, verificando o que ocorre no histograma, média e desvio padrão de `ma`. Atente para o fato de que valores de `n` maiores que 1000 podem tornar o processo custoso computacionalmente.

4.3.3 Amostra representativa

Ouve-se frequentemente o argumento de que uma boa amostra é aquela que é representativa. Indagado sobre a definição de uma amostra representativa, a resposta mais comum é algo como: "aquela que é uma micro representação do universo". Mas para se ter certeza de que uma amostra seja uma micro representação do universo para uma dada característica de interesse, deve-se conhecer o comportamento dessa mesma característica da população. Então, o conhecimento da população seria tão grande que tonar-se-ia desnecessária

a coleta da amostra.
(Bolfarine and Bussab, 2005, p. 14)

4.3.4 Tipos de amostras

Critério	Procedimento de Seleção	
	probabilístico	não probabilístico
objetivo	amostras probabilísticas	amostras criteriosas
subjetivo	amostras quase-aleatórias	amostras intencionais

Figure 4.1: Tipos de amostras segundo (Bolfarine and Bussab, 2005) e (Jessen, 1978).

Procedimentos probabilísticos objetivos são mais bem aceitos academicamente, ainda que na prática nem sempre possam ser executados. Quando isso ocorre, podem-se considerar procedimentos que sejam possíveis de serem executados.

4.4 Principais técnicas de amostragem

4.4.1 Amostragem Aleatória Simples

Amostragem Aleatória Simples (AAS) é o método mais básico de seleção de amostras, sendo referência para todos os demais planos amostrais. A partir de uma lista completa das N unidades elementares da população seleciona-se cada unidade amostral com igual probabilidade, de tal forma que a cada sorteio os elementos tenham a mesma chance de serem escolhidos. A necessidade de uma lista completa da população para realizar uma AAS pode eventualmente ser um limitante na aplicação deste tipo de metodologia, pois na prática nem sempre é possível obter tal listagem. Os planos amostrais A e B discutidos nos Exemplos 4.16, 4.19 e 4.20 são caracterizados como AAS.

AAS sem reposição - AASs

Caso a unidade sorteada seja retirada da população e seja realizado um novo sorteio, é dito que procedeu-se com uma AAS *restrita* ou *sem reposição*, indicado por AASs.

Exemplo 4.26. (AASs) De uma urna com N cartões numerados de 1 a N sorteia-se um ao acaso, com probabilidade $1/N$. O cartão sorteado é deixado de fora da urna e realiza-se um novo sorteio, onde cada um dos $N-1$ cartões restantes tem probabilidade $1/(N-1)$ de ser retirado. Este procedimento é repetido até que se concluam todos os n sorteios desejados. Este é um processo de AAS *sem* reposição.

Exercício 4.6. Defina as probabilidades de sorteio do Exemplo 4.26 considerando $n = 3$ sorteios.

Exercício 4.7. Calcule as probabilidades de sorteio do Exemplo 4.26 considerando $n = 3$ sorteios e $N = 10$.

Exemplo 4.27. (Mega-Sena da Caixa Econômica Federal) No R pode-se sortear uma amostra sem reposição para tentar a sorte no jogo da Caixa Econômica Federal. Note que $N = 60$, $n = 6$.

```
set.seed(1234) # Fixando a geração pseudo-aleatória
sort(sample(1:60, size = 6, replace = F)) # Apostando na Mega-Sena da CEF via AASs
```

```
## [1] 16 22 28 37 44 58
```

Exercício 4.8. Leia a documentação das funções utilizadas no Exemplo 4.27 fazendo `?set.seed`,
`?sort` e `?sample`.

AAS com reposição - AASc

Caso a unidade sorteada tenha a chance de participar novamente da amostra, o procedimento é chamado AAS *irrestrita* ou *com reposição*, indicado por AASc.

Exemplo 4.28. (AASc) De uma urna com N cartões numerados de 1 a N sorteia-se um ao acaso, com probabilidade $1/N$. O cartão sorteado é recolocado na urna e realiza-se um novo sorteio. Este procedimento é repetido até que se concluam todos os n sorteios desejados. Este é um processo de AAS *com* reposição.

Exercício 4.9. Defina as probabilidades de sorteio do Exemplo 4.28 considerando $n = 3$ sorteios.

Exercício 4.10. Calcule as probabilidades de sorteio do Exemplo 4.28 considerando $n = 3$ sorteios com $N = 10$.

4.4.2 Amostragem Sistemática

Considere uma população de N unidades elementares numeradas de 1 a N. Para selecionar uma amostra de n observações sorteia-se aleatoriamente uma

das primeiras $k = N/n$ unidades, digamos a, com probabilidade $1/k$ selecionando sistematicamente as próximas observações a cada k unidades. Matematicamente,

$$a, a + k, a + 2k, \dots, a + (n - 1)k.$$

Exemplo 4.29. Seja uma população com $N = 100$, da qual deseja-se retirar uma amostra sistemática de tamanho $n = 5$. Neste caso $k = 100/5 = 20$, então sortearmos aleatoriamente um número entre 1 e 20 com probabilidade $1/20$. Se o número sorteado for $a = 4$, a amostra sistemática então é definida como

$$4, 4 + 20, 4 + 2 \times 20, 4 + 3 \times 20, 4 + 4 \times 20 = 4, 24, 44, 64, 84.$$

```
N <- 100 # Tamanho da população
n <- 5 # Tamanho da amostra
(k <- N/n) # Tamanho do salto
```

```
## [1] 20
```

```
set.seed(1) # Fixando a geração pseudo-aleatória
(a <- sample(1:k, size = 1)) # Amostra de tamanho 1 com probabilidade 1/k
```

```
## [1] 4
```

```
for(i in 1:n){print(a+(i-1)*k)} # Apresentando a amostra de tamanho n
```

```
## [1] 4
## [1] 24
## [1] 44
## [1] 64
## [1] 84
```

Exercício 4.11. Considere o código do Exemplo 4.29.
(a) Rode o código repetidas vezes retirando a linha `set.seed(1)`. O que você observa?
(b) Refaça o exercício considerando outros valores de N e n, tais que $n < N$.

4.4.3 Amostragem Estratificada

Segundo (Bolfarine and Bussab, 2005, p. 93),

> Amostragem estratificada consiste na divisão de uma população em grupos (estratos) segundo alguma(s) característica(s) conhecida(s) na populaçãao sob estudo, e de cada um desses estratos são selecionadas amostras em proporções convenientes.

O objetivo deste tipo de amostragem é que o pesquisador possa se valer de estruturas pré-existentes de maneira a melhorar as inferências, reduzindo sua variabilidade.

4.5 Cálculo do tamanho da amostra

O *cálculo do tamanho de amostra* é baseado em uma série de premissas assumidas pelo pesquisador. Os valores sugeridos pelos diversos métodos de cálculo de tamanho de amostra devem ser considerados apenas como uma referência, dada a arbitrariedade das medidas utilizadas em sua obtenção. Tempo e custo são dois limitantes que devem ser levados em conta, podendo se sobrepor aos cálculos de tamanho de amostra.
A seguir serão apresentados casos bastante simples, mas suficientes para ilustrar os princípios utilizados. Para mais funcionalidades recomenda-se o pacote `pwr` (Champely, 2020) do R e o software G*Power (Franz Faul and Buchner, 2007), (Franz Faul and Lang, 2009). Para uma abordagem mais teórica recomenda-se (Chow et al., 2007).

4.5.1 Média

Uma forma de estimar o tamanho da amostra no caso da inferência para a media universal μ é considerar a margem de erro da Equação (5.2) e isolar n na forma

$$n = \left\lceil \left(\frac{z\sigma}{\varepsilon} \right)^2 \right\rceil. \tag{4.7}$$

O operador $\lceil x \rceil$ indica a função *teto* de x, i.e., indica o primeiro inteiro acima de x.

Exercício 4.12. Obtenha o resultado da Equação (4.7) a partir da margem de erro da Equação (5.2).

Exemplo 4.30. (Tamanho da amostra para a média) Deseja-se obter o tamanho de amostra para estimar a média de altura dos alunos da PUCRS. Considera-se um intervalo de confiança de $1 - \alpha = 95\%$, com margem de erro de $\varepsilon = 3$ cm. De estudos anteriores, admite-se $\sigma = 15$ cm. Considerando a Equação (4.7), sabe-se da tabela da distribuição normal padrão que $z = 1.96$, assim

$$n = \left\lceil \left(\frac{1.96 \times 15}{3} \right)^2 \right\rceil = \lceil 96.04 \rceil = 97. \tag{4.8}$$

```
# Equação (3.8)
n_m <- function(z,sigma,e) {
  exato <- (z*sigma/e)^2
  teto <- ceiling(exato)
  return(list(exato=exato,
              teto=teto))
}
n_m(1.96,15,3)
```

```
## $exato
## [1] 96
##
## $teto
## [1] 97
```

```
n_m(1.96,15,3)$exato
```

```
## [1] 96
```

```
n_m(1.96,15,3)$teto
```

```
## [1] 97
```

4.5.2 Proporção

Uma forma de estimar o tamanho da amostra no caso da inferência para a proporção universal π é considerar a margem de erro da Equação (5.4) e isolar n na forma

$$n = \left\lceil \frac{z^2 p(1-p)}{\varepsilon^2} \right\rceil. \tag{4.9}$$

Em certos casos existe informação disponível sobre a proporção, mas quando não há qualquer conhecimento a respeito desta medida considera-se $p = \frac{1}{2}$, ponto no qual $p(1-p)$ atinge seu máximo.

Exercício 4.13. Obtenha o resultado da Equação (4.9) a partir da margem de erro da Equação (5.4).

Exercício 4.14. Verifique que $p(1-p)$ atinge seu máximo quando $p = \frac{1}{2}$.

Exemplo 4.31. (Tamanho da amostra para a proporção) Em uma pesquisa eleitoral deseja-se calcular o tamanho de amostra aproximado para que a margem de erro seja de $\varepsilon = 2\%$ com confiança de $1 - \alpha = 95\%$. Considerando a Equação (4.9), sabe-se da tabela da distribuição normal padrão que $z = 1.96 \approx 2$, e que $p(1-p)$ atinge seu máximo quando $p = \frac{1}{2}$. Assim,

$$n \approx \left\lceil \frac{2^2 \times \frac{1}{2} \times (1 - \frac{1}{2})}{\varepsilon^2} \right\rceil = \left\lceil \frac{1}{\varepsilon^2} \right\rceil \tag{4.10}$$

Logo, um IC para a proporção com $\alpha = 5\%$ para uma margem de erro de $\varepsilon = 2\%$ pode ser calculado com um tamanho de amostra de aproximadamente

$$n \approx \left\lceil \frac{1}{0.02^2} \right\rceil = 2500.$$

```r
# Equação (3.11)
n_p <- function(e) {
  exato <- 1/e^2
  teto <- ceiling(exato)
  return(list(exato=exato,
              teto=teto))
}
n_p(0.02)
```

```
## $exato
## [1] 2500
##
## $teto
## [1] 2500
```

```r
n_p(0.02)$exato
```

```
## [1] 2500
```

```r
n_p(0.02)$teto
```

```
## [1] 2500
```

Exercício 4.15. Teste a função `n_p` do Exemplo 4.31 com diferentes valores de margem de erro. Faça um gráfico para analisar a variação do tamanho da amostra à medida que ε aumenta.

4.6 Para saber mais

O material Amostragem: Teoria e Prática Usando R, gentilmente disponibilizado pelos professores Pedro Luis do Nascimento Silva, Zélia Magalhães Bianchini e Antonio José Ribeiro Dias, é uma fonte muito rica para informações adicionais sobre este tópico. Está disponível ainda o livro Análise de Dados Amostrais Complexos, também do professor Pedro Silva em parceria com o professor Djalma Pessoa.
O professor Pedro também compartilhou o vídeo e os slides da apresentação *Combinando amostras para aprimorar estimativas – aventuras na amostragem não probabilística*, apresentado em 18 de outubro de 2020 no *VII Encontro Baiano de Estatística*.

Chapter 5

Inferência Clássica

Inferência é o procedimento que visa atualizar a informação sobre parâmetros a partir dos dados amostrais. Para (DeGroot and Schervish, 2012, p. 378), 'inferência estatística é um procedimento que produz uma declaração probabilística a respeito de uma ou mais partes de um modelo estatístico'[1]. Conclusões obtidas a partir dos dados embasam a *lógica indutiva*, i.e., aquela que parte do caso particular para o caso geral e que se opõe à *lógica dedutiva* que vai do caso geral para o particular. O princípio indutivo se enquadra na noção de um decisor, a partir de uma amostra (caso particular), *inferir* sobre parâmetros universais/populacionais.

(Berger, 1985, p.16) e (Paulino et al., 2003, p. 6) indicam que os procedimentos baseados no *paradigma clássico* baseiam-se em alguns princípios, tais como máxima verossimilhança, não viés[2], variância mínima, mínimos quadrados, consistência, suficiência e eficiência. Os clássicos consideram que existe um parâmetro θ desconhecido para o qual não se atribui probabilidades. A amostra é obtida *aleatoriamente* de um universo de interesse, sendo uma das tantas - se não infinitas - possíveis amostras. (Berger, 1985, p.26) aponta que tal princípio coloca os clássicos como *incondicionalistas*, pois pondera-se sobre todos os dados possíveis e não condicionado ao que foi observado.

A abordagem clássica possui três tipos de estimação: *Pontual* (ou *por ponto*), por *Intervalo de Confiança* (IC/ICo) e por *Teste de Hipóteses* (TH), detalhadas a seguir.

[1] '*A statistical inference is a procedure that produces a probabilistic statement about some or all parts of a statistical model.*'

[2] Definição 5.3.

5.1 Estimação Pontual

Na *estimação pontual* utiliza-se uma estatística, calculada a partir de um *estimador* como *estimativa* (*pontual*) de um certo parâmetro, conforme Definição 5.1 e 5.2. Em outras palavras, é utilizado um único valor amostral (ponto) para estimar θ, simbolizado por $\hat{\theta}$ e lido como *teta chapéu*.

Definição 5.1. Um *estimador* $\hat{\theta}(x) \equiv \hat{\theta}$ é uma função que tem por objetivo inferir sobre um parâmetro $\theta(X) \equiv \theta$.

Definição 5.2. Uma *estimativa* é um particular valor obtido da aplicação dos dados amostrais em um estimador.

Exemplo 5.1. A média amostral $\bar{x}$ é um estimador pontual para a média universal μ (Eq. (4.2)). É dado pela Equação (2.9).

5.1.1 Estimadores e suas propriedades

Definição 5.3. Um estimador é dito *não viesado* ou *não viciado* segundo um plano amostral λ se

$$E_\lambda \left[\hat{\theta}\right] = \theta. \tag{5.1}$$

Média amostral $\bar{x}$

A média amostral do Exemplo (2.9) é um estimador não viesado da média universal μ segundo o plano amostral AAS, com ou sem reposição. Isto ocorre pelo fato de a esperança ser linear, portanto a dependência entre as observações não interfere no resultado.

Exemplo 5.2. Sejam as variáveis aleatórias $X_1, X_2, \ldots, X_n$ independentes identicamente distribuídas (iid) com $E(X_i) = \mu$ e um plano amostral do tipo AAS,

onde por simplicidade será considerada a equivalência $E_{AAS} \equiv E$.

$$\begin{aligned} E[\bar{x}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[x_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \frac{1}{n} n\mu \\ E[\bar{x}] &= \mu. \end{aligned}$$

Exemplo 5.3. A média universal da variável `idade` do Exemplo 4.4 é dada por

$$\mu = \frac{24 + 32 + 49}{3} = \frac{105}{3} = 35.$$

Do Exemplo 4.19 pode-se verificar que a média (esperança) das médias amostrais considerando o plano AASc é igual a μ, i.e.,

$$E[h(x)] = E[\bar{x}] = \frac{24.0 + 28.0 + 36.5 + 28.0 + 32.0 + 40.5 + 36.5 + 40.5 + 49.0}{9} = \frac{315}{9} = 35.$$

```
X <- c(24,32,49)
mean(X)
```

```
## [1] 35
```

Do Exemplo 4.22 tem-se o vetor `mxc <- c(24.0,28.0,36.5,28.0,32.0,40.5,36.5,40.5,49.0)`.

```
mean(mxc)
```

```
## [1] 35
```

Exercício 5.1. Verifique no plano amostral AASs do Exemplo 4.20 que $E[\bar{x}] = \mu$.

Proporção amostral p

A *proporção amostral* é um estimador não viesado da proporção universal π (Eq. (4.3)) segundo o plano amostral AAS, com ou sem reposição. Pode-se definir este estimador por

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

Exemplo 5.4. (Estimativa pontual da proporção) Suponha que deseja-se calcular a estimativa pontual para a 'proporção de fumantes da PUCRS', denotada por π. A característica de interesse, ou sucesso, é o entrevistado ser 'fumante', para o qual associa-se $x = 1$; desta forma, o fracasso é o entrevistado ser 'não fumante', para o qual associa-se $x = 0$. Em uma amostra de $n = 125$ frequentadores da universidade, observaram-se $\sum_{i=1}^{n} x_i = 25$ fumantes. A estimativa pontual de π é dada por

$$\widehat{\pi} = \frac{25}{125} = 0.2 = 20\%.$$

Variância amostral s^2

A variância amostral é um estimador não viesado da variância universal σ^2 segundo o plano amostral AAS com reposição.

Exemplo 5.5. Sejam as variáveis aleatórias $X_1, X_2, \ldots, X_n$ independentes identicamente distribuídas (iid) com $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $E(X_i^2) = \sigma^2 + \mu^2$ e um plano amostral do tipo AASc, onde por simplicidade será considerada a equivalência $E_{AASc} \equiv E$.

$$\begin{aligned}
E\left[s^2\right] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{y})^2\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left[x_i^2\right] - E\left[n\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left[x_i^2\right] - nE\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right] \\
&= \frac{(n-1)\sigma^2}{n-1} \\
E\left[s^2\right] &= \sigma^2
\end{aligned}$$

Exercício 5.2. Verifique no plano amostral AASc do Exemplo 4.19 se $E_{AASc}\left[s^2\right] = \sigma^2$.

Exercício 5.3. Verifique no plano amostral AASs do Exemplo 4.20 se $E_{AASs}\left[s^2\right] = \sigma^2$.

5.2 (Estimação por) Intervalo de Confiança

5.2.1 Média

σ conhecido

O IC para a média universal com σ conhecido é dado pela expressão

$$IC\left[\mu, 1-\alpha\right] = \bar{x} \mp z\frac{\sigma}{\sqrt{n}} = \left[\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right], \tag{5.2}$$

onde $1-\alpha$ é a confiança do intervalo, $\bar{x}$ é a média amostral, σ é o desvio padrão universal conhecido, n é o tamanho da amostra e $z = z_{\frac{\alpha}{2}}$ é o quantil da distribuição normal padrão que acumula $\frac{\alpha}{2}$ de probabilidade.

Exemplo 5.6. (IC para μ com σ conhecido) Considere uma amostra de $n = 10$ mulheres, das quais observou-se a variável X: 'altura'. Suponha que $X \sim \mathcal{N}(\mu, 0.05)$, i.e., a variável 'altura das mulheres' tem distribuição normal com média μ desconhecida e desvio padrão universal conhecido $\sigma = 0.05$. Da tabela da distribuição normal sabe-se que os quantis ± 1.96 limitam uma área de aproximadamente 95%, portanto $z = 1.96$. Se a média da amostra é $\bar{x}_{10} = 1.63$, o IC de $1-\alpha = 95\%$ é

$$IC\left[\mu, 95\%\right] = 1.63 \mp 1.96\frac{0.05}{\sqrt{10}} \approx 1.63 \mp 0.03 \approx \left[1.60, 1.66\right].$$

A margem de erro é de aproximadamente 0.03 ou 3 cm.

```
n <- 10
m <- 1.63
sigma <- 0.05                    # 'sigma' universal
z <- abs(qnorm(0.025))           # |-1.959964|
(e <- z*sigma/sqrt(n))           # Margem de erro
```

```
## [1] 0.031
```

```
(LImusig <- m - e)               # Limite Inferior
```

```
## [1] 1.6
```

```
(LSmusig <- m + e)               # Limite Superior
```

```
## [1] 1.66
```

```
# Princípio de relatório automático
cat('O IC 95% para a média é [',
    round(LImusig,2), ',',
    round(LSmusig,2), '].')
```

```
## O IC 95% para a média é [ 1.6 , 1.66 ].
```

σ desconhecido

O IC para a média universal com σ desconhecido é dado pela expressão

$$IC\left[\mu, 1-\alpha\right] = \bar{x} \mp t\frac{s}{\sqrt{n}} = \left[\bar{x} - t\frac{s}{\sqrt{n}}, \bar{x} + t\frac{s}{\sqrt{n}}\right], \tag{5.3}$$

onde $1-\alpha$ é a confiança do intervalo, $\bar{x}$ é a média amostral, s é o desvio padrão amostral, n é o tamanho da amostra e $t = t_{n-1,\frac{\alpha}{2}}$ é o quantil da distribuição t com $n-1$ graus de liberdade que acumula $1-\frac{\alpha}{2}$ de probabilidade. Note a troca de σ por s, implicando na penalidade de utilizar t no lugar de z.

Exemplo 5.7. (IC para μ com σ desconhecido) Assim como no Exemplo 5.6, considere uma amostra de $n = 10$ mulheres das quais observou-se a variável X: 'altura'. Suponha que $X \sim \mathcal{N}(\mu, \sigma)$, i.e., a variável 'altura das mulheres' tem distribuição normal com média μ e desvio padrão universal σ, ambos desconhecidos. Da tabela da distribuição t com $10-1 = 9$ graus de liberdade, sabe-se que os quantis ± 2.262 limitam uma área de aproximadamente 95%, portanto $t = 2.262$. Se da amostra calculou-se média de $\bar{x}_{10} = 1.63$ e desvio padrão de $s = 0.05$, o IC de $1-\alpha = 95\%$ é

$$IC\left[\mu, 95\%\right] = 1.63 \mp 2.262\frac{0.05}{\sqrt{10}} \approx 1.63 \mp 0.04 \approx \left[1.59, 1.67\right].$$

A margem de erro é de aproximadamente 0.04 ou 4 cm, maior que a margem de erro 0.03 quando assume-se σ conhecido pois $z = 1.96 < 2.262 = t$.

```
n <- 10
m <- 1.63
s <- 0.05                          # 's' amostral
t <- abs(qt(0.025, n-1))           # |-2.2621572|
(e <- t*s/sqrt(n))                 # Margem de erro
```

```
## [1] 0.0358
```

```
(LImus <- m - e)                   # Limite Inferior
```

```
## [1] 1.59
```

```
(LSmus <- m + e)                   # Limite Superior
```

```
## [1] 1.67
```

```
# Princípio de relatório automático
cat('O IC 95% para a média é [',
    round(LImus,2), ',',
    round(LSmus,2), '].')
```

```
## O IC 95% para a média é [ 1.59 , 1.67 ].
```

5.2.2 Proporção

O IC para a proporção populacional π é dado pela expressão

$$IC\left[\pi, 1-\alpha\right] = p \mp z\sqrt{\frac{p(1-p)}{n}} = \left[p - z\sqrt{\frac{p(1-p)}{n}}, p + z\sqrt{\frac{p(1-p)}{n}}\right] \quad (5.4)$$

onde $1-\alpha$ é a confiança do intervalo, p é a proporção amostral, n é o tamanho da amostra e $z = z_{\frac{\alpha}{2}}$ é o quantil da distribuição normal padrão que acumula $\frac{\alpha}{2}$ de probabilidade.

Exemplo 5.8. (IC para π) Considere novamente os dados do Exemplo 5.4, onde deseja-se calcular o IC para a proporção de fumantes da PUCRS. Sabe-se que $\hat{\pi} = p = 25/125 = 0.2$, $n = 125$ e $z = 1.96$. O IC de $1-\alpha = 95\%$ é

$$IC\left[\pi, 95\%\right] = 0.2 \mp 1.96\sqrt{\frac{0.2\,(1-0.2)}{125}} \approx 0.2 \mp 0.07 = [0.13, 0.27] = [13\%, 27\%]\,.$$

A margem de erro é de aproximadamente $0.07 = 7\%$, Note a diferença de precisão entre a tabela, consultando a probabilidade 0.0250 correspondente a $z = -1.96$, e o valor calculado com a função `qnorm`.

```
n <- 125
p <- 25/n
z <- abs(qnorm(0.025))          # |-1.959964|
(e <- z*sqrt(p*(1-p)/n))        # Margem de erro
```

```
## [1] 0.0701
```

```
(LIpi <- p - e)                 # Limite Inferior
```

```
## [1] 0.13
```

```
(LSpi <- p + e)                 # Limite Superior
```

```
## [1] 0.27
```

```
# Princípio de relatório automático
cat('O IC 95% para a proporção é [',
    round(LIpi,2), ',',
    round(LSpi,2), '].')
```

```
## O IC 95% para a proporção é [ 0.13 , 0.27 ].
```

Exercício 5.4. Acesse o material Estatística Clássica no RStudio e resolva os exercícios extras 1 ao 9 das páginas 99 e 100. Observe o Apêndice B com as respostas dos exercícios, mas só após tentar resolvê-los.

5.3 (Estimação por) Teste de Hipóteses

Seja um parâmetro θ pertencente a um *espaço paramétrico* Θ, i.e., o conjunto de todos os possíveis valores de θ. Considere uma partição tal que $\Theta = \Theta_0 \cup \Theta_1$ e $\Theta_0 \cap \Theta_1 = \emptyset$. Um *teste de hipóteses* é uma regra de decisão que permite decidir, à luz das informações disponíveis, se é mais verossímil admitir $\theta \in \Theta_0$ ou $\theta \in \Theta_1$. A hipótese que envolve Θ_0 é chamada *hipótese nula*, e a que envolve Θ_1 é a *hipótese alternativa*. Tais hipóteses podem ser escritas na estrutura de *Neyman-Pearson*, na forma

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases} .$$

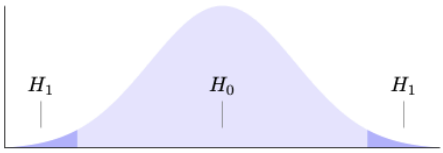
Usualmente nos procedimentos de testes de hipóteses admite-se inicialmente que H_0 seja verdadeira, dito *sob* H_0. Por este motivo a hipótese nula sempre deve conter a igualdade, o que indicado no quadro abaixo. Note que não há uma 'regra da hipótese nula', a indicação está colocada desta forma apenas por motivos didáticos.

Regra da Hipótese Nula

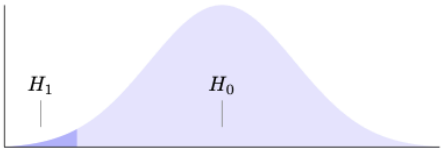
A igualdade sempre está em H_0.

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases} \text{ ou } \begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \text{ ou } \begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

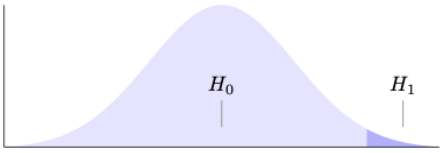
As definições do quadro acima implicam em três tipos de hipótese alternativa, conforme figura a seguir. A hipótese *bilateral* é a típica hipótese de equilíbrio, geralmente utilizada quando não há definição prévia sobre a direção da hipótese, tal como no caso de decidir se uma moeda deve ou não ser considerada equilibrada. A hipótese *unilateral inferior* é uma hipótese que indica um piso de referência, tal como no caso de decidir sobre a eficácia mínima de um tratamento (maior, melhor). A hipótese *unilateral superior* é uma hipótese que indica um teto de referência, tal como no caso de decidir sobre a uma ação dependente de uma taxa máxima de mortalidade (menor, melhor).



(a) Hipótese bilateral ($H_1 : \theta \neq \theta_0$)



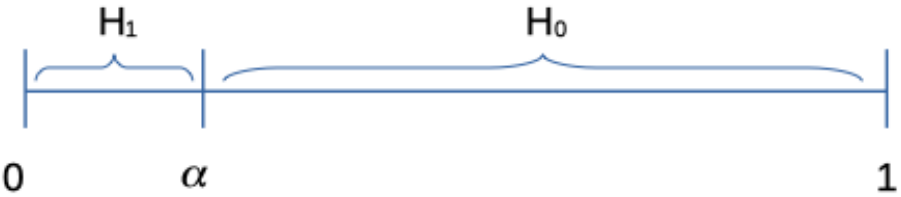
(b) Hipótese unilateral inferior ($H_1 : \theta < \theta_0$)



(c) Hipótese unilateral superior ($H_1 : \theta > \theta_0$)

Exercício 5.5. Para cada item abaixo, indique as hipóteses sendo testadas.
a. A companhia de transporte afirma que, em média, o intervalo entre sucessivos ônibus de uma determinada linha é de 15 minutos. Uma associação de usuários de transportes coletivos acha que a pontualidade é muito importante, e deseja testar a afirmação da companhia.
b. Os amortecedores de automóveis que circulam em cidades duram pelo menos 100 mil quilômetros em média, segundo a informação de algumas oficinas especializadas. O proprietário de uma locadora de veículos deseja testar esta afirmação.
c. Um veterinário afirma ter obtido um ganho médio diário de pelo menos 3 litros de leite por vaca com uma nova composição de ração. Um pecuarista acredita que o ganho não é tão grande assim.
d. Algumas garrafas de cerveja declaram em seus rótulos conter 600mL. Os órgãos de fiscalização desejam avaliar se uma fábrica deve ou não ser autuada por engarrafar cervejas com uma quantidade menor que o indicado no rótulo.
e. O dado de um cassino parece estar viciado, saindo o valor 1 com uma frequência muito grande.
f. Um fabricante afirma que a sua vacina previne pelo menos 80% dos casos de uma doença. Um grupo de médicos desconfia que a vacina não é tão eficiente assim.

A partir da premissa de que H_0 é verdadeira, compara(m)-se o(s) valor(es) descrito(s) nesta hipótese com os dados da amostra através de uma medida chamada *estatística de teste* ou *quantidade pivotal*. Caso a estatística do teste indique uma pequena distância entre o(s) valor(es) de H_0 e a estatística, *admite-se* ou *não se rejeita* H_0; caso a distância seja grande, *rejeita-se* H_0. As distâncias que fazem admitir ou rejeitar H_0 são avaliadas em termos probabilísticos, indicadas nos gráficos acima respectivamente pelas regiões claras e escuras. A divisão destas regiões é dada por *valores críticos*, quantis das distribuições associadas às estatísticas de teste que serão detalhadas a seguir.



É possível ainda considerar formas mais precisas de avaliar as distâncias probabilisticas das estatísticas de teste do que simplesmente indicando 'acima' ou 'abaixo' de um valor crítico. Pelo paradigma clássico, atribui-se uma medida que varia entre 0 e 1, chamada *p-value*, *valor-p* ou *nível descritivo amostral*. Este medida possui múltiplas definições e ainda é bastante discutida na literatura. Não é raro ser mal interpretada, portanto será considerada como um *grau de evidência em favor de* H_0. Rejeita-se H_0 se este grau de evidência for baixo, inferior a um valor de referência chamado *nível de significância* e representado por α; caso contrário, admite-se ou não se rejeita H_0. Este valor de significância está associado à probabilidade de *erro do tipo I*, ou o caso em que erramos ao rejeitar uma hipótese H_0 verdadeira. Tal valor é *arbitrário*, ou seja, deve ser definido pelo dono do problema ao estipular o quanto admite de probabilidade máxima de erro do tipo I. Existem valores de significância típicos, usualmente 10%, 5%, 1%, e 0.1%. Devido a um exemplo dado por (Fisher, 1925)[3], o valor de 5% tornou-se uma referência para o valor de α, ainda que existam propostas mais elaboradas e melhor embasadas na teoria Estatística. Destacam-se os trabalhos de (Gannon et al., 2019), (Pereira and Stern, 1999) e (Pereira and Wechsler, 1993).

Associação com intervalos de confiança

Os testes de hipóteses possuem as mesmas características e propriedades dos seus respectivos intervalos de confiança. Desta forma, apresenta-se um breve exemplo abordando a equivalência entre os TH e os IC para a proporção universal π.

Exemplo 5.9. (TH $\equiv$ IC) Suponha uma moeda com probabilidade de face cara $Pr(H) = \pi$. Em princípio não sabemos o valor de π, e pode ser interessante considerar duas configurações:

$$\begin{cases} H_0 : \text{a moeda é equilibrada} \\ H_1 : \text{a moeda não é equilibrada} \end{cases} \equiv \begin{cases} H_0 : \pi = 0.5 \\ H_1 : \pi \neq 0.5 \end{cases}.$$

Sob H_0 (i.e., supondo H_0 verdadeira), espera-se observar 'cara' em 50% dos resultados, com alguma variação em torno de 50%. Considerando a Equação

[3] *"The value for which P=.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant."* Ronald Aylmer Fisher na primeira edição do seu livro *Statistical Methods For Research Workers*, de 1925.

(5.4) pode-se obter a margem de erro esperada para esta oscilação em função do tamanho da amostra n para, digamos, 95% dos casos:

$$IC\left[\pi, 95\%\right] = 0.5 \mp 1.96\sqrt{\frac{0.5\left(1-0.5\right)}{n}} = 0.5 \mp \frac{0.98}{\sqrt{n}}.$$

Assim, ao realizar $n = 100$ lançamentos e observar uma frequência de caras no intervalo

$$IC\left[\pi, 95\%\right] = 0.5 \mp \frac{0.98}{\sqrt{100}} = [0.402, 0.598] = [40.2\%, 59.8\%]$$

pode-se considerar a moeda equilibrada com $1 - \alpha = 95\%$ de confiança. Caso a frequência seja inferior a 40.2% ou superior a 59.8%, há indícios de que a moeda seja desequilibrada, também com 95% de confiança. Pela terminologia dos testes de hipóteses, não se rejeita H_0 com $\alpha = 5\%$.
Se $n = 25$,

$$IC\left[\pi, 95\%\right] = 0.5 \mp \frac{0.98}{\sqrt{25}} = [0.304, 0.696] = [30.4\%, 69.6\%]$$

e obtém-se um intervalo maior se comparado a $n = 100$, i.e., menos preciso para a mesma confiança de 95%. Como exercício, use a função `ic` para definir outros valores para n e teste este resultado em uma moeda.

```
# IC95% sob H0: \pi=0.5
ic <- function(n){
  cat('[', 0.5-.98/sqrt(n), ',',
           0.5+.98/sqrt(n), ']')
}
ic(100)
```

```
## [ 0.402 , 0.598 ]
```

```
ic(25)
```

```
## [ 0.304 , 0.696 ]
```

5.3.1 Testes Paramétricos Univariados

TESTE 1 - Teste z para média de uma amostra

Hipótese avaliada
Uma amostra de n sujeitos (ou objetos) vem de uma população de média μ igual a um valor especificado μ_0?

Suposições
S1. O desvio padrão universal σ é conhecido;
S2. A amostra foi selecionada aleatoriamente da população que representa;

S3. A distribuição dos dados na população que a amostra representa é normal. Pelo Teorema Central do Limite, tal suposição torna-se menos importante à medida que o tamanho da amostra aumenta.

Testes relacionados
TESTE 13 - Teste dos postos sinalizados de Wilcoxon para uma amostra

Estatística do teste
Sob $H_0 : \mu = \mu_0$, $H_0 : \mu \geq \mu_0$ ou $H_0 : \mu \leq \mu_0$,

$$z_{teste} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1). \tag{5.5}$$

Valor-p
Sob $H_0 : \mu = \mu_0$,

$$\text{Valor-p} = 2Pr(Z \leq -|z_{teste}|). \tag{5.6}$$

Sob $H_0 : \mu \geq \mu_0$,

$$\text{Valor-p} = Pr(Z \leq z_{teste}). \tag{5.7}$$

Sob $H_0 : \mu \leq \mu_0$,

$$\text{Valor-p} = Pr(Z \geq z_{teste}). \tag{5.8}$$

Exemplo 5.10. É desejado testar se a média de altura dos alunos da PUCRS pode ser considerada *maior* do que 167 cm. A hipótese é portanto unilateral superior na forma $H_0 : \mu \leq 167$ *vs* $H_1 : \mu > 167$. Estudos anteriores indicam que a variável X: 'altura dos alunos da PUCRS' tem distribuição normal de média desconhecida (motivo da realização do teste de hipóteses para μ) e desvio padrão $\sigma = 14$, indicado por $X \sim \mathcal{N}(\mu, 14)$. De uma amostra aleatória com $n = 25$ pessoas obteve-se $\bar{x}_{25} = 172$. Assim, sob H_0 a estatística do teste pode ser calculada da seguinte maneira:

$$z_{teste} = \frac{172 - 167}{14/\sqrt{25}} \approx 1.786 \approx 1.79.$$

Se utilizarmos $\alpha = 0.05$ (unilateral superior), $z_{crítico} = 1.64$. Como a estatística de teste extrapola o valor crítico, i.e., $1.79 > 1.64$, rejeita-se H_0.

Decisão Estatística: Rejeita-se H_0 com $\alpha = 5\%$ pois $1.79 > 1.64$.
Conclusão Experimental: A amostra sugere que a média de altura dos alunos da PUCRS deve ser maior do que 167 cm.

Exemplo 5.11. No Exemplo 5.10 é possível calcular o valor p associado à estatística de teste $z_{teste} \approx 1.79$. Por ser um teste unilateral superior, basta obter a probabilidade de encontrar um valor tão ou mais extremo que z_{teste}

conforme Equação (5.7). Pela tabela de normal padrão (com precisão inferior à do computador)

$$\text{Valor-p} = Pr(Z \geq 1.79) = 1 - Pr(Z < 1.79) = 1 - 0.9633 = 0.0367.$$

Utilizando $\alpha = 0.05$ unilateral decide-se novamente pela rejeição de H_0 uma vez que o valor p é inferior ao nível de significância, i.e., $0.0367 < 0.05$. A decisão realizada desta maneira deve sempre ser a mesma quando compara-se a estatística do teste com o(s) valor(es) crítico(s).

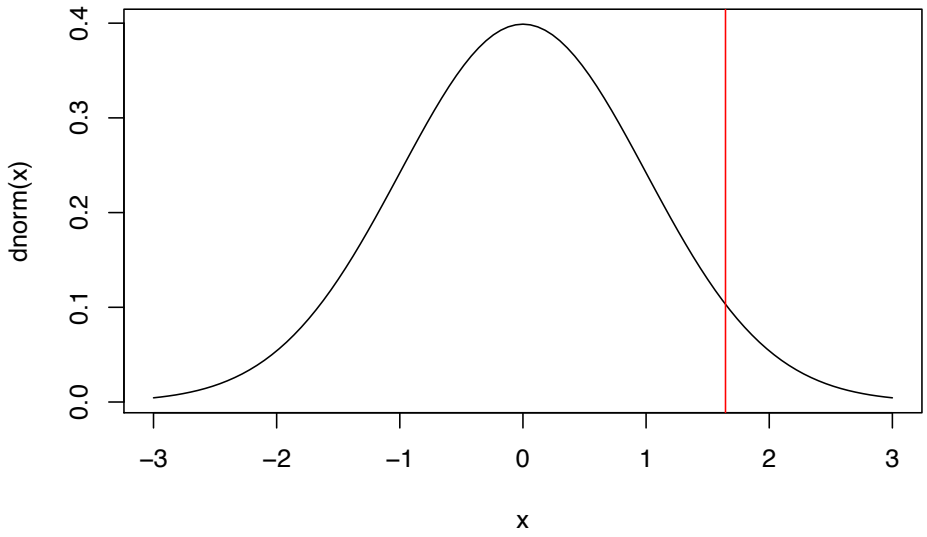
Decisão Estatística: Rejeita-se H_0 com $\alpha = 5\%$ pois $0.0367 < 0.05$.
Conclusão Experimental: A amostra sugere que a média de altura dos alunos da PUCRS deve ser maior do que 167 cm.

Exemplo 5.12. Realizando os Exemplos 5.10 e 5.11 no R.

```
# Definindo os valores indicados no enunciado
mu0 <- 167
n <- 25
x_bar <- 172
sigma <- 14
(zt <- (x_bar-mu0)/(sigma/sqrt(n))) # estatística do teste, note a maior precisão
```

```
## [1] 1.79
```

```
curve(dnorm(x), -3, 3) # gráfico da normal padrão
abline(v = qnorm(.95), col = 'red') # valor crítico de 1.64
```



```r
1-pnorm(zt) # p-value mais preciso por conta de zt e pnorm
```

```
## [1] 0.0371
```

TESTE 2 - Teste t para média de uma amostra

Hipótese avaliada
Uma amostra de n sujeitos (ou objetos) vem de uma população de média μ igual a um valor especificado μ_0?

Suposições
S1. A amostra foi selecionada aleatoriamente da população que representa;
S2. A distribuição dos dados na população que a amostra representa é normal.

Testes relacionados
TESTE 14 - Teste dos postos sinalizados de Wilcoxon para uma amostra

Estatística do teste
Sob $H_0 : \mu = \mu_0$,

$$t_{teste} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(gl), \tag{5.9}$$

onde $gl = n - 1$ indica os *graus de liberdade* que definem a distribuição t.

Valor-p
Sob $H_0 : \mu = \mu_0$,

$$\text{Valor-p} = 2Pr(T \leq -|t_{teste}|). \tag{5.10}$$

Sob $H_0 : \mu \geq \mu_0$,

$$\text{Valor-p} = Pr(T \leq t_{teste}). \tag{5.11}$$

Sob $H_0 : \mu \leq \mu_0$,

$$\text{Valor-p} = Pr(T \geq t_{teste}). \tag{5.12}$$

Exemplo 5.13. É desejado testar se a média de altura dos alunos da PUCRS pode ser considerada *maior* do que 167 cm. A hipótese é portanto unilateral superior na forma $H_0 : \mu \leq 167$ *vs* $H_1 : \mu > 167$. Estudos anteriores indicam que a variável X: 'altura dos alunos da PUCRS' tem distribuição normal de média e desvio padrão desconhecidos, indicado por $X \sim \mathcal{N}(\mu, \sigma)$. De uma amostra aleatória com $n = 25$ pessoas obteve-se $\bar{x}_{25} = 172$ e $s_{25} = 14$. Assim, sob H_0 a estatística do teste pode ser calculada da seguinte maneira:

$$t_{teste} = \frac{172 - 167}{14/\sqrt{25}} \approx 1.786.$$

Se utilizarmos $\alpha = 0.05$ (unilateral superior), $t_{crítico} = 1.711$, considerando $gl = 24$ graus de liberdade. Como a estatística de teste extrapola o valor crítico, i.e., $1.786 > 1.711$, rejeita-se H_0.

Decisão Estatística: Rejeita-se H_0 com $\alpha = 5\%$ pois $1.786 > 1.711$.
Conclusão Experimental: A amostra sugere que a média de altura dos alunos da PUCRS deve ser maior do que 167 cm.

Exemplo 5.14. No Exemplo 5.13 é possível obter um intervalo para o valor p associado à estatística de teste $t_{teste} \approx 1.786$. Por ser um teste unilateral superior, deve-se obter a probabilidade de encontrar um valor tão ou mais extremo que t_{teste} conforme Equação (5.11). Pela tabela de t com $gl = 24$ graus de liberdade obtém-se $Pr(t > 1.711) = 0.05$ e $Pr(t > 2.064) = 0.025$. Dada a limitação de precisão da tabela t, pode-se apenas concluir que $0.025 < Pr(t > 1.786) < 0.05$. Utilizando $\alpha = 0.05$ unilateral decide-se novamente pela rejeição de H_0 uma vez que o valor p é inferior ao nível de significância, i.e., $Pr(t > 1.786) < 0.05$. A decisão realizada desta maneira deve sempre ser a mesma quando compara-se a estatística do teste com o(s) valor(es) crítico(s).

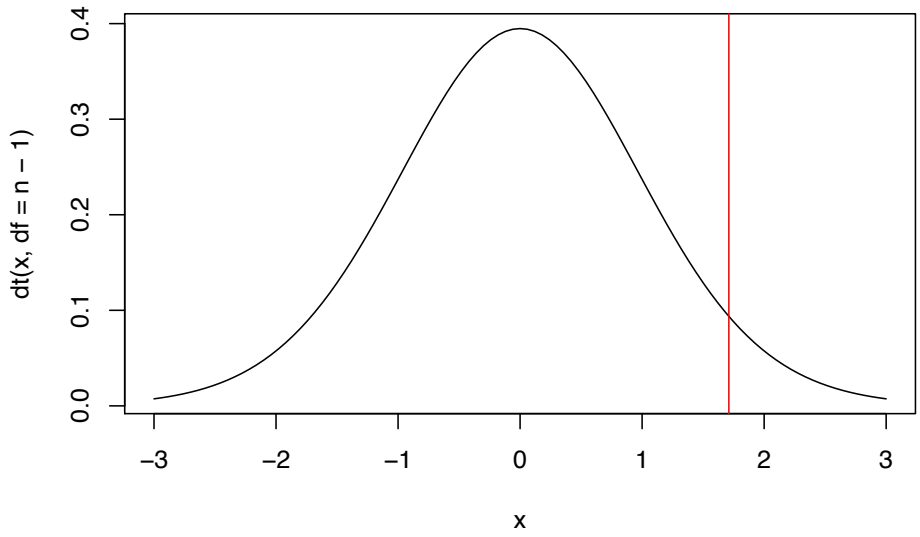
Decisão Estatística: Rejeita-se H_0 com $\alpha = 5\%$ pois $Pr(t > 1.786) < 0.05$.
Conclusão Experimental: A amostra sugere que a média de altura dos alunos da PUCRS deve ser maior do que 167 cm.

Exemplo 5.15. Realizando os Exemplos 5.13 e 5.14 no R.

```
# Definindo os valores indicados no enunciado
mu0 <- 167
n <- 25
x_bar <- 172
s <- 14
(tt <- (x_bar-mu0)/(s/sqrt(n))) # estatística do teste, note a maior precisão
```

```
## [1] 1.79
```

```
curve(dt(x, df = n-1), -3, 3) # gráfico da t com gl=25-1=24
abline(v = qt(.95, df = n-1), col = 'red') # valor crítico de  1.711
```




```
1-pt(tt, df = n-1) # p-value mais preciso por conta de tt e pt
```

```
## [1] 0.0434
```

TESTE 3 - Testes para proporção de uma amostra, binomial (exato) e normal (assintótico)

Hipótese avaliada
Em uma população composta de duas categorias, a proporção π de observações em uma das categorias é igual a um valor específico π_0?

Suposições
S1. Cada observação pode ser classificada em *sucesso* ou *fracasso*;
S2. Cada uma das n observações (condicionalmente) independentes é selecionada aleatoriamente de uma população;
S3. A probabilidade de sucesso π se mantém constante a cada observação.

Estatística do teste (assintótico)
Sob $H_0 : \pi = \pi_0$,

$$z_{teste} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim \mathcal{N}(0,1). \tag{5.13}$$

Valor-p (assintótico)
Sob $H_0 : \pi = \pi_0$, vide Equação (5.6).

Sob $H_0 : \pi \geq \pi_0$, vide Equação (5.7).

Sob $H_0 : \pi \leq \pi_0$, vide Equação (5.8).

Valor-p (exato)
Seja X o número de sucessos em n ensaios de Bernoulli. Sob $H_0 : \pi = \pi_0$ ocorre

que $X \sim \mathcal{B}(n, \pi_0)$, se $x > \frac{n}{2}$ e $I = \{0, 1, ..., n-x, x, ..., n\}$,

$$\text{Valor-p} = Pr(n - x \geq X \geq x | \pi = \pi_0) = \sum_{i \in I} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}, \tag{5.14}$$

se $x < \frac{n}{2}$ e $I = \{0, 1, ..., x, n-x, ..., n\}$,

$$\text{Valor-p} = Pr(x \geq X \geq n - x | \pi = \pi_0) = \sum_{i \in I} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}, \tag{5.15}$$

e Valor-p $= 1$ se $x = \frac{n}{2}$.

Sob $H_0 : \pi \leq \pi_0$,

$$\text{Valor-p} = Pr(X \geq x | \pi = \pi_0) = \sum_{i=x}^{n} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i} \tag{5.16}$$

Sob $H_0 : \pi \geq \pi_0$,

$$\text{Valor-p} = Pr(X \leq x | \pi = \pi_0) = \sum_{i=0}^{x} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i} \tag{5.17}$$

Exemplo 5.16. Suponha que deseja-se testar π, a proporção de caras em uma moeda, na forma $H_0 : \pi \leq 0.5$ *vs* $H_1 : \pi > 0.5$. Para isso a moeda é lançada $n = 12$ vezes, onde se observam $x = 9$ caras e $n - x = 12 - 9 = 3$ coroas. Sabe-se que $p = \frac{9}{12} = \frac{3}{4} = 0.75$. Considerando a abordagem assintótica, sob H_0

$$z_{teste} = \frac{0.75 - 0.5}{\sqrt{0.5(1 - 0.5)/12}} \approx 1.73.$$

Se utilizarmos $\alpha = 0.05$ (unilateral superior), $z_{crítico} = 1.64$. Como a estatística de teste extrapola o valor crítico, i.e., $1.73 > 1.64$, rejeita-se H_0.

Decisão Estatística: Rejeita-se H_0 com $\alpha = 5\%$ pois $1.73 > 1.64$.
Conclusão Experimental: A amostra sugere que a proporção de caras da moeda deve ser considerada maior que 0.5.

Exemplo 5.17. Considere novamente os dados do Exemplo 5.16. Pela Equação (5.8) utilizando a tabela de normal padrão (com precisão inferior à do computador),

$$\text{Valor-p} = Pr(Z \geq 1.73) = 1 - Pr(Z < 1.73) = 1 - 0.9582 = 0.0418.$$

Utilizando $\alpha = 0.05$ unilateral decide-se novamente pela rejeição de H_0 uma vez que o valor p é inferior ao nível de significância, i.e., $0.0418 < 0.05$. A

decisão realizada desta maneira deve sempre ser a mesma quando compara-se a estatística do teste com o(s) valor(es) crítico(s).

Decisão Estatística: Rejeita-se H_0 com $\alpha = 5\%$ pois $0.0418 < 0.05$.
Conclusão Experimental: A amostra sugere que a proporção de caras da moeda deve ser considerada maior que 0.5.

Exemplo 5.18. Realizando os Exemplos 5.16 e 5.17 no R.

```
n <- 12
x <- 9
(p <- x/n)
```

```
## [1] 0.75
```

```
pi0 <- 0.5
(zt <- (p-pi0)/sqrt(pi0*(1-pi0)/n))
```

```
## [1] 1.73
```

```
1-pnorm(zt) # p-value mais preciso por conta de zt e pnorm
```

```
## [1] 0.0416
```

```
# usando a função prop.test, sem a correção de Yates
prop.test(x, n, pi0, alternative = 'greater', correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability pi0
## X-squared = 3, df = 1, p-value = 0.04
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.513 1.000
## sample estimates:
##    p
## 0.75
```

Exemplo 5.19. Considere novamente os dados do Exemplo 5.16. O teste exato pode ser realizado considerando que sob $H_0 : \pi \leq 0.5$, o número de caras (sucessos) X tem distribuição binomial de parâmetros $n = 12$ e $\pi = 0.5$, i.e., $X \sim \mathcal{B}(12, 0.5)$. Assim, o valor-p exato resulta em

$$Pr\left(X \geq 9 | \pi = 0.5\right) = \left[\binom{12}{9} + \binom{12}{10} + \binom{12}{11} + \binom{12}{12}\right] \times 0.5^{12} \approx 0.0730.$$

Note a diferença do valor exato em comparação ao assintótico.
Decisão Estatística: não se rejeita H_0 com $\alpha = 5\%$ pois $0.0730 > 0.05$.

Conclusão Experimental: a amostra sugere que a proporção de caras da moeda pode ser considerada menor ou igual a 0.5.

Exemplo 5.20. Realizando o Exemplo 5.19 no R.

```
# manualmente
n <- 12
x <- 9
pi0 <- 0.5
p9 <- dbinom(9,n,pi0)
p10 <- dbinom(10,n,pi0)
p11 <- dbinom(11,n,pi0)
p12 <- dbinom(12,n,pi0)
p9+p10+p11+p12 # valor-p
```

```
## [1] 0.073
```

```
# usando a função binom.test
binom.test(x, n, pi0, alternative = 'greater')
```

```
##
##  Exact binomial test
##
## data:  x and n
## number of successes = 9, number of trials = 12, p-value = 0.07
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.473 1.000
## sample estimates:
## probability of success
##                   0.75
```

```
# usando a função prop.test (assintótico mas com correção de continuidade de Yates)
prop.test(x, n, pi0, alternative = 'greater')
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability pi0
## X-squared = 2, df = 1, p-value = 0.07
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.471 1.000
## sample estimates:
##    p
## 0.75
```

Exercício 5.6. Refaça os Exemplos 5.16, 5.17 e 5.19 considerando $H_0 : \pi = 0.6$ *vs* $H_1 : \pi \neq 0.6$.

Exercício 5.7. Uma rádio do estado anunciou que 90% dos hotéis da Serra Gaúcha estariam lotados no final de semana do dia dos pais. A estação aconselhou os ouvintes a fazerem reserva antecipada para se hospedar na Serra nestes dias. No sábado à noite uma amostra de 58 hotéis revelou que 49 diziam 'sem vagas'. Qual é a sua reação à afirmação da rádio, depois de ver a evidência da amostra? Use 5% de nível de significância.

Exercício 5.8. Você é responsável por avaliar a qualidade de um grande lote de peças de segunda mão adquiridas pela sua empresa. O fabricante afirma haver no máximo 10% de peças defeituosas, e você decide investigar. Para isso você retira uma amostra de 50 peças, das quais 9 são defeituosas. Qual a sua opinião sobre o lote adquirido, considerando níveis de significância de 1%, 5% e 10%? Defina as hipóteses, apresentando a Decisão Estatística e a Conclusão Experimental.

TESTE 4 - Teste qui-quadrado para a variância populacional de uma amostra

Hipótese avaliada
Uma amostra de n sujeitos (ou objetos) vem de uma população na qual a variância σ^2 é igual a um valor especificado σ_0^2?

Suposições
S1. A amostra foi selecionada aleatoriamente da população que representa;
S2. A distribuição dos dados na população que a amostra representa é normal.

Estatística do teste
Sob $H_0 : \sigma^2 = \sigma_0^2$,

$$\chi^2_{teste} = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(gl), \tag{5.18}$$

onde $gl = n - 1$ indica os *graus de liberdade* que definem a distribuição χ^2.

Valor-p
Sob $H_0 : \sigma^2 = \sigma_0^2$,

$$\text{Valor-p} = 2Pr(\chi^2 \leq \chi^2_{teste}). \tag{5.19}$$

Sob $H_0 : \sigma^2 \geq \sigma_0^2$,

$$\text{Valor-p} = Pr(\chi^2 \leq \chi^2_{teste}). \tag{5.20}$$

Sob $H_0 : \sigma^2 \leq \sigma_0^2$,

$$\text{Valor-p} = Pr(\chi^2 \geq \chi^2_{teste}). \tag{5.21}$$

Exemplo 5.21. Deseja-se testar se a variância de uma variável com distribuição normal pode ser considerada igual a 5, i.e., $H_0 : \sigma^2 = 5$ *vs* $H_0 : \sigma^2 \neq 5$. Para isso observa-se uma amostra de tamanho $n = 41$, de onde se calcula uma variância amostral de $s^2 \approx 3.196876$. Sob H_0

$$\chi^2_{teste} = \frac{(41-1) \times 3.196876}{5} \approx 25.58.$$

Considerando $\alpha = 0.05$ (bilateral) e a tabela qui-quadrado com $gl = 41-1 = 40$, $\chi^2_{crítico1} = 24.43$ e $\chi^2_{crítico2} = 59.34$. Como a estatística de teste não extrapola os valores críticos, i.e., $24.43 < 25.58 < 59.34$, não se rejeita H_0.

Decisão Estatística: Não se rejeita H_0 com $\alpha = 0.05$ pois $24.43 < 25.58 < 59.34$.
Conclusão Experimental: A amostra sugere que a variância da referida variável pode ser considerada igual a 5.

Exemplo 5.22. Realizando o Exemplo 5.21 no R.

```
# Definindo os valores indicados no enunciado
sigma2_0 <- 5
n <- 41
set.seed(123); x <- rnorm(n, mean = 0, sd = 2)
(s2 <- var(x))
```

```
## [1] 3.2
```

```
(quit <- ((n-1)*s2)/sigma2_0) # estatística do teste, note a maior precisão
```

```
## [1] 25.6
```

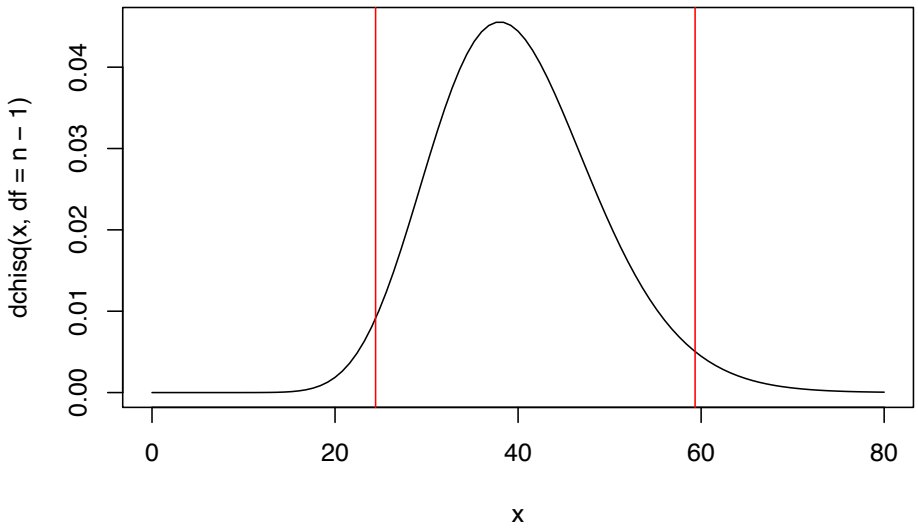
```
curve(dchisq(x, df = n-1), 0, 80) # gráfico da qui^2 com gl=41-1=40
(qui_cr1 <- qchisq(.025, df = n-1)) # valor crítico 1
```

```
## [1] 24.4
```

```
(qui_cr2 <- qchisq(.975, df = n-1)) # valor crítico 2
```

```
## [1] 59.3
```

```
abline(v = c(qui_cr1, qui_cr2), col = 'red') # valores críticos
```




```
2*pchisq(quit, df = n-1) # p-value bilateral, H1: sigma^2  5, Eq. (3.24)
```

```
## [1] 0.0743
```

```
pchisq(quit, df = n-1) # p-value unilateral inferior, H1: sigma^2 < 5, Eq. (3.25)
```

```
## [1] 0.0372
```

```
1-pchisq(quit, df = n-1) # p-value unilateral superior, H1: sigma^2 > 5, Eq. (3.26)
```

```
## [1] 0.963
```

```
# Via pacote DescTools
library(DescTools)
VarTest(x, sigma.squared = sigma2_0, alternative = 'two.sided')
```

```
##
##  One Sample Chi-Square test on variance
##
## data:  x
## X-squared = 26, df = 40, p-value = 0.07
## alternative hypothesis: true variance is not equal to 5
## 95 percent confidence interval:
##  2.15 5.23
## sample estimates:
## variance of x
##           3.2
```

```
VarTest(x, sigma.squared = sigma2_0, alternative = 'less')
```

```
##
##  One Sample Chi-Square test on variance
```

```
##
## data:  x
## X-squared = 26, df = 40, p-value = 0.04
## alternative hypothesis: true variance is less than 5
## 95 percent confidence interval:
##  0.00 4.82
## sample estimates:
## variance of x
##           3.2
```

```
VarTest(x, sigma.squared = sigma2_0, alternative = 'greater')
```

```
##
##  One Sample Chi-Square test on variance
##
## data:  x
## X-squared = 26, df = 40, p-value = 1
## alternative hypothesis: true variance is greater than 5
## 95 percent confidence interval:
##  2.29  Inf
## sample estimates:
## variance of x
##           3.2
```

Exercício 5.9. Suponha que o comprimento de peças em uma fábrica, simbolizado por X, tenha distribuição normal de média e variância desconhecidas, anotado por $X \sim \mathcal{N}(\mu, \sigma)$. A especificação indica média de 140cm, e desvio padrão de 7cm. Se em uma amostra de 64 peças foi observada uma média de $\bar{x} = 138$cm e um desvio padrão de $s = 12$cm, realize os testes de hipóteses apropriados para avaliar se as especificações estão sendo cumpridas.

TESTE 5 - Teste qui-quadrado de aderência de uma amostra

Hipótese avaliada
No universo representado por uma amostra, há diferença entre as frequências esperadas e observadas?

Suposições
S1. Os dados avaliados consistem em uma amostra aleatória de n observações (condicionalmente) independentes;
S2. Os dados representam frequências de k categorias mutuamente exclusivas.

Estatística do teste
Sob $H_0 : \pi_1 = \pi_1^0, \pi_2 = \pi_2^0, \ldots, \pi_k = \pi_k^0$,

$$\chi^2_{teste} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(gl), \tag{5.22}$$

onde $E_i = n\pi_i^0$, k é o número de categorias e $gl = k - 1$ indica os *graus de liberdade* que definem a distribuição χ^2.

Valor-p
Sob $H_0 : \pi_1 = \pi_1^0, \pi_2 = \pi_2^0, \ldots, \pi_k = \pi_k^0$,

$$\text{Valor-p} = Pr(\chi^2 \geq \chi^2_{teste}). \tag{5.23}$$

Exemplo 5.23. (Adaptado de (Sheskin, 2011, p.278) - Teste qui-quadrado de aderência balanceado) Um dado é lançado 120 vezes, a fim de determinar se pode ou não ser considerado equilibrado. Os valores observados estão apresentados conforme tabela abaixo, e $E_i = 120 \times \frac{1}{6} = 20, i \in \{1, 2, 3, 4, 5, 6\}$.

Face (i)	1	2	3	4	5	6
O_i	20	14	18	17	22	29
E_i	20	20	20	20	20	20

Sob $H_0 : \pi_1 = \frac{1}{6}, \pi_2 = \frac{1}{6}, \pi_3 = \frac{1}{6}, \pi_4 = \frac{1}{6}, \pi_5 = \frac{1}{6}, \pi_6 = \frac{1}{6}$,

$$\chi^2_{teste} = \frac{(20-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(29-20)^2}{20} = 6.7.$$

Considerando $\alpha = 0.05$ (unilateral superior, que é sempre o caso deste teste) e a tabela qui-quadrado com $gl = 6 - 1 = 5$, $\chi^2_{crítico} = 11.07$. Como a estatística de teste não extrapola os valores críticos, i.e., $6.7 < 11.07$, não se rejeita H_0. Considerando a Equação (5.23) e a função `pchisq`,

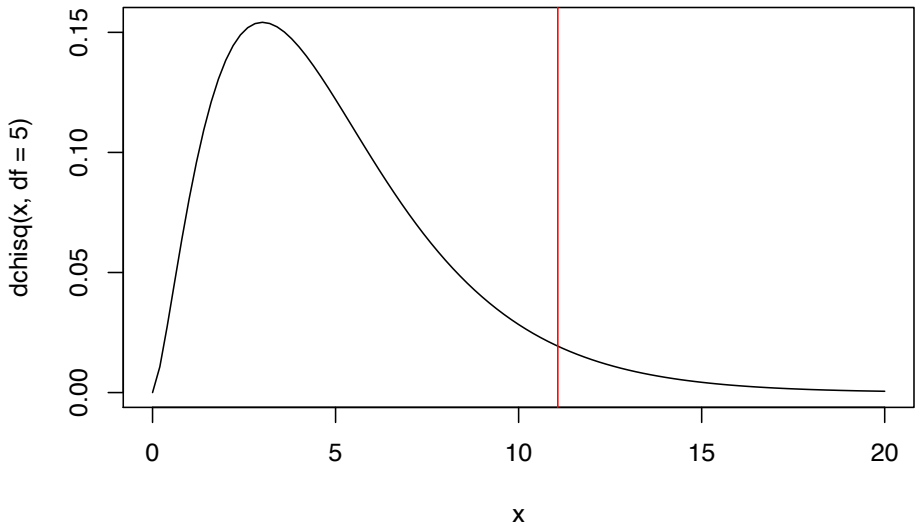
$$\text{Valor-p} = Pr(\chi^2 \geq 6.7) \approx 0.2439.$$

Decisão Estatística: Não se rejeita H_0 com $\alpha = 0.05$ pois $6.7 < 11.07$ ou $0.2439 > 0.05$.
Conclusão Experimental: A amostra sugere que o dado deve ser equilibrado.

```
curve(dchisq(x, df=5), 0, 20) # gráfico da qui^2 com gl=6-1=5
(qui_cr <- qchisq(.95, df=5)) # valor crítico
```

```
## [1] 11.1
```

```
abline(v = qui_cr, col='red') # valor crítico no gráfico
```




```r
o <- c(20,14,18,17,22,29)      # Observados
n <- sum(o)                    # Tamanho da amostra
p <- rep(1/6,6)                # Distribuição uniforme (dado equilibriado)
e <- n*p                       # Valores esperados se o dado for equilibrado
k <- length(o)                 # Número de categorias
(qui <- sum((o-e)^2/e))        # Equação (3.25)
```

```
## [1] 6.7
```

```r
1-pchisq(qui,k-1)              # p-value
```

```
## [1] 0.244
```

```r
chisq.test(o)                  # Pela função 'chisq.test'
```

```
##
##  Chi-squared test for given probabilities
##
## data:  o
## X-squared = 7, df = 5, p-value = 0.2
```

Exemplo 5.24. (Teste qui-quadrado de aderência desbalanceado) Gregor Mendel conduziu experimentos sobre hereditariedade em ervilhas. Em suma, as ervilhas podiam ser redondas (R) ou enrugadas (E), amarelas (A) ou verdes (V). Portanto, existem quatro combinações possíveis: RA, RV, EA, EV. Se sua teoria estivesse correta, as ervilhas seriam observadas na proporção de 9:3:3:1. Se o resultado do experimento produziu os seguintes dados observados, pode-se avaliar se há indícios da proporção considerada[4].

[4]Ainda que (Fisher, 1936) tenha posto dúvida sobre o trabalho de Mendel ao criar a noção do *Paradoxo Mendeliano*, ou 'bom demais para ser verdade'. Tal consideração tem bases

i	RA (1)	RV (2)	EA (3)	EV (4)	n
O_i	315	108	101	32	**556**
E_i	312.75	104.25	104.25	34.75	**556**

Sob $H_0 : \pi_1 = \frac{9}{16}, \pi_2 = \frac{3}{16}, \pi_3 = \frac{3}{16}, \pi_4 = \frac{1}{16}$,

$$\chi^2_{teste} = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} \approx 0.47.$$

Considerando $\alpha = 0.05$ (unilateral superior, que é sempre o caso deste teste) e a tabela qui-quadrado com $gl = 4-1 = 3$, $\chi^2_{crítico} = 7.81$. Como a estatística de teste não extrapola os valores críticos, i.e., $0.47 < 7.81$, não se rejeita H_0. Considerando a Equação (5.23) e a função `pchisq`,

$$\text{Valor-p} = Pr(\chi^2 \geq 0.47) \approx 0.9254.$$

Decisão Estatística: Não se rejeita H_0 com $\alpha = 0.05$ pois $0.47 < 7.81$ ou $0.9254 > 0.05$.
Conclusão Experimental: A amostra sugere que a proporção das ervilhas deve ser 9:3:3:1.

```
o <- c(315,108,101,32)       # Observados
(n <- sum(o))                # Tamanho da amostra
```

```
## [1] 556
```

```
(k <- length(o))             # Número de categorias
```

```
## [1] 4
```

```
p <- c(9/16,3/16,3/16,1/16) # Proporção 9:3:3:1
(e <- n*p)                   # Valores esperados se a prop. for 9:3:3:1
```

```
## [1] 312.8 104.2 104.2  34.8
```

```
(qui <- sum((o-e)^2/e))      # Estatística do teste
```

```
## [1] 0.47
```

```
1-pchisq(qui,k-1)            # Valor p
```

```
## [1] 0.925
```

eugenistas e incondicionalistas, calcada sob uma ótica ultrapassada assumida por Fisher e seus mentores, Karl Pearson e Francis Galton e já desmentida por acadêmicos como (Novitski, 2004) e (Hartl and Fairbanks, 2007).

```
chisq.test(o,p=p)           # Pela função 'chisq.test'
```

```
##
##  Chi-squared test for given probabilities
##
## data:  o
## X-squared = 0.5, df = 3, p-value = 0.9
```

5.3.2 Testes Paramétricos Bivariados

TESTE 6 - Teste F (de Hartley) de igualdade de variâncias

Hipótese avaliada
A variância do universo 1 é igual à variância do universo 2.

Suposições
S1. Os tamanhos de amostra $n1$ e $n2$ são similares.

Estatística do teste
Sob $H_0 : \sigma_1 = \sigma_2$,

$$F_{max} = \frac{s^2_{max}}{s^2_{min}} \sim \mathcal{F}(n_{max} - 1, n_{min} - 1), \tag{5.24}$$

onde s^2_{max} e s^2_{min} são respectivamente a maior e menor variância amostral, e n_{max} e n_{min} correspondem respectivamente ao tamanho de amostra associado à amostra de maior e menor variância.

Valor-p
Sob $H_0 : \sigma_1 = \sigma_2$,

$$\text{Valor-p} = 2Pr(F \geq F_{max}). \tag{5.25}$$

Sob $H_0 : \sigma_1 \geq \sigma_2$,

$$\text{Valor-p} = Pr(F \geq F_{max}). \tag{5.26}$$

Sob $H_0 : \sigma_1 \leq \sigma_2$,

$$\text{Valor-p} = Pr(F < F_{max}). \tag{5.27}$$

```
x <- 1:10
y <- c(7:20)
nx <- length(x)
ny <- length(y)
(Fmax <- var(y)/var(x))
```

```
## [1] 1.91
```

```
2*(1-pf(Fmax,ny-1,nx-1)) # H_0:  _1 =  _2
```

```
## [1] 0.334
```

```r
1-pf(Fmax,ny-1,nx-1) # H_0: _1   _2
```

```
## [1] 0.167
```

```r
pf(Fmax,ny-1,nx-1) # H_0: _1   _2
```

```
## [1] 0.833
```

```r
# Utilizando a função var.test
var.test(x,y) # H_0: _1 = _2
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.5, num df = 9, denom df = 13, p-value = 0.3
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.158 2.007
## sample estimates:
## ratio of variances
##              0.524
```

```r
var.test(x,y, alternative = 'less') # H_0: _1   _2
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.5, num df = 9, denom df = 13, p-value = 0.2
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.0 1.6
## sample estimates:
## ratio of variances
##              0.524
```

```r
var.test(x,y, alternative = 'greater') # H_0: _1   _2
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.5, num df = 9, denom df = 13, p-value = 0.8
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.193   Inf
## sample estimates:
```

```
## ratio of variances
##              0.524
```

TESTE 7 - Teste z para médias de duas amostras independentes

Hipótese avaliada
Duas amostras independentes representam duas populações com valores médios diferentes?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. A distribuição dos dados na população subjacente de cada amostra é normal;
S3. (*Homogeneidade de variâncias*) A variância da população representada pela amostra 1 é igual à variância da população representada pela amostra 2 ($\sigma_1^2 = \sigma_2^2$).

Testes relacionados
TESTE 16 - Teste dos postos de Mann-Whitney para duas amostras independentes

Estatística do teste
Sob $H_0 : \mu_1 - \mu_2 = \Delta_0$,

$$z_{teste} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1), \tag{5.28}$$

onde n_1 e n_2 são os tamanhos das amostras, $\bar{x}_1$ e $\bar{x}_2$ representam as médias amostrais e σ_1^2 e σ_2^2 são as variâncias universais dos universos 1 e 2.

Valor-p
Sob $H_0 : \mu_1 - \mu_2 = \Delta_0$,

$$\text{Valor-p} = 2Pr(Z \leq -|z_{teste}|). \tag{5.29}$$

Sob $H_0 : \mu_1 - \mu_2 \geq \Delta_0$,

$$\text{Valor-p} = Pr(Z \leq z_{teste}). \tag{5.30}$$

Sob $H_0 : \mu_1 - \mu_2 \leq \Delta_0$,

$$\text{Valor-p} = Pr(Z \geq z_{teste}). \tag{5.31}$$

Exemplo 5.25.

```
x <- 1:10
y <- c(7:20)
nx <- length(x)
```

```
ny <- length(y)
mx <- mean(x)
my <- mean(y)
sigmax2 <- var(x)*(nx-1)/nx
sigmay2 <- var(y)*(ny-1)/ny
(zt <- (mx-my)/sqrt(sigmax2/nx+sigmay2/ny))
```

```
## [1] -5.68
```

```
2*pnorm(-abs(zt)) # H_0: μ_1-μ_2 = 0
```

```
## [1] 1.37e-08
```

```
pnorm(zt)         # H_0: μ_1-μ_2  0
```

```
## [1] 6.85e-09
```

```
1-pnorm(zt)       # H_0: μ_1-μ_2  0
```

```
## [1] 1
```

TESTE 8 - Teste t para médias de duas amostras independentes

Hipótese avaliada
Duas amostras independentes representam duas populações com valores médios diferentes?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. A distribuição dos dados na população subjacente de cada amostra é normal.

Testes relacionados
TESTE 16 - Teste dos postos de Mann-Whitney para amostras independentes

Estatística do teste
Sob $H_0 : \mu_1 - \mu_2 = \Delta_0$ e $\sigma_1 = \sigma_2$,

$$t_{teste} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\left[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right]\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}} \sim t(gl). \tag{5.32}$$

Sob $H_0 : \mu_1 - \mu_2 = \Delta_0$ e $\sigma_1 \neq \sigma_2$,

$$t_{teste} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t(gl), \tag{5.33}$$

onde n_1 e n_2 são os tamanhos das amostras, $\bar{x}_1$ e $\bar{x}_2$ representam as médias amostrais e s_1^2 e s_2^2 são as variâncias amostrais dos universos 1 e 2. Se as

variâncias forem admitidas iguais ($\sigma_1 = \sigma_2$), os graus de liberdade são calculados utilizando a expressão

$$gl = n_1 + n_2 - 2.$$

No caso de as variâncias serem admitidas diferentes ($\sigma_1 \neq \sigma_2$), calculam-se os graus de liberdade com a abordagem de Welch, dados por

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Valor-p
Sob $H_0 : \mu_1 - \mu_2 = \Delta_0$,

$$\text{Valor-p} = 2Pr(t \leq -|t_{teste}|). \tag{5.34}$$

Sob $H_0 : \mu_1 - \mu_2 \geq \Delta_0$,

$$\text{Valor-p} = Pr(t \leq t_{teste}). \tag{5.35}$$

Sob $H_0 : \mu_1 - \mu_2 \leq \Delta_0$,

$$\text{Valor-p} = Pr(t \geq t_{teste}). \tag{5.36}$$

Exemplo 5.26.

```
x <- 1:10
y <- c(7:20)
nx <- length(x)
ny <- length(y)
mx <- mean(x)
my <- mean(y)
sx2 <- var(x)
sy2 <- var(y)
sp2 <- ((nx-1)*sx2 + (ny-1)*sy2)/(nx+ny-2)
# dando uma olhada nas variâncias
var.test(x,y)
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.5, num df = 9, denom df = 13, p-value = 0.3
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.158 2.007
```

```
## sample estimates:
## ratio of variances
##              0.524
```

```
# estatística do teste para variâncias iguais
(tt_igual <- (mx-my)/sqrt(sp2*(1/nx+1/ny)))
```

```
## [1] -5.15
```

```
# estatística do teste para variâncias diferentes
(tt_welch <- (mx-my)/sqrt(sx2/nx+sy2/ny))
```

```
## [1] -5.43
```

```
# graus de liberdade para variâncias iguais
(gl_igual <- nx+ny-2)
```

```
## [1] 22
```

```
# graus de liberdade para variâncias diferentes
(gl_welch <- (sx2/nx+sy2/ny)^2/((sx2/nx)^2/(nx-1) + (sy2/ny)^2/(ny-1)))
```

```
## [1] 22
```

```
# H_0: μ_1-μ_2 = 0, variâncias iguais
2*pt(-abs(tt_igual), gl_igual)
```

```
## [1] 3.69e-05
```

```
# H_0: μ_1-μ_2  0, variâncias iguais
pt(tt_igual, gl_igual)
```

```
## [1] 1.85e-05
```

```
# H_0: μ_1-μ_2  0, variâncias iguais
1-pt(tt_igual, gl_igual)
```

```
## [1] 1
```

```
# H_0: μ_1-μ_2 = 0, variâncias diferentes
2*pt(-abs(tt_welch), gl_welch)
```

```
## [1] 1.86e-05
```

```
# H_0: μ_1-μ_2  0, variâncias diferentes
pt(tt_welch, gl_welch)
```

```
## [1] 9.28e-06
```

```
# H_0: μ_1-μ_2  0, variâncias diferentes
1-pt(tt_welch, gl_welch)
```

```
## [1] 1
```

```r
# via t.test (facilita uma barbaridade!)
# H_0: μ_1-μ_2 = 0,  _1 =  _2
t.test(1:10, y = c(7:20), var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  1:10 and c(7:20)
## t = -5, df = 22, p-value = 4e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.22  -4.78
## sample estimates:
## mean of x mean of y
##       5.5      13.5
```

```r
# H_0: μ_1-μ_2   0,  _1 =  _2
t.test(1:10, y = c(7:20), var.equal = T, alternative = 'less')
```

```
##
##  Two Sample t-test
##
## data:  1:10 and c(7:20)
## t = -5, df = 22, p-value = 2e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##   -Inf -5.33
## sample estimates:
## mean of x mean of y
##       5.5      13.5
```

```r
# H_0: μ_1-μ_2   0,  _1 =  _2
t.test(1:10, y = c(7:20), var.equal = T, alternative = 'greater')
```

```
##
##  Two Sample t-test
##
## data:  1:10 and c(7:20)
## t = -5, df = 22, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -10.7   Inf
## sample estimates:
## mean of x mean of y
##       5.5      13.5
```

```
# H_0: μ_1-μ_2 = 0, _1   _2
t.test(1:10, y = c(7:20), var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  1:10 and c(7:20)
## t = -5, df = 22, p-value = 2e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.05  -4.95
## sample estimates:
## mean of x mean of y
##       5.5      13.5
```

```
# H_0: μ_1-μ_2  0, _1   _2
t.test(1:10, y = c(7:20), var.equal = F, alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  1:10 and c(7:20)
## t = -5, df = 22, p-value = 9e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##   -Inf -5.47
## sample estimates:
## mean of x mean of y
##       5.5      13.5
```

```
# H_0: μ_1-μ_2  0, _1   _2
t.test(1:10, y = c(7:20), var.equal = F, alternative = 'greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  1:10 and c(7:20)
## t = -5, df = 22, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -10.5   Inf
## sample estimates:
## mean of x mean of y
##       5.5      13.5
```

TESTE 9 - Teste t para médias de duas amostras dependentes/pareadas

Hipótese avaliada
Duas amostras dependentes representam duas populações com médias diferentes?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. A distribuição dos dados na população subjacente de cada amostra é normal;
S3. (*Homogeneidade de variâncias*) A variância da população representada pela amostra 1 é igual à variância da população representada pela amostra 2 ($\sigma_1^2 = \sigma_2^2$).

Testes relacionados
TESTE 17 - Teste dos postos de Wilcoxon para amostras dependentes/pareadas.

Estatística do teste
Sob $H_0 : \mu_1 - \mu_2 = \Delta_0$,

$$t_{teste} = \frac{\bar{D} - \Delta_0}{s_{\bar{D}}/\sqrt{n}} \sim t(gl), \tag{5.37}$$

onde

$$\bar{D} = \frac{\sum D}{n},$$

$D = x_1 - x_2$ e

$$s_{\bar{D}} = \sqrt{\left(\frac{\sum D^2}{n} - \bar{D}^2\right)\left(\frac{n}{n-1}\right)}. \tag{5.38}$$

Valor-p
Sob $H_0 : \mu_1 - \mu_2 = \Delta_0$,

$$\text{Valor-p} = 2Pr(t \leq -|t_{teste}|). \tag{5.39}$$

Sob $H_0 : \mu_1 - \mu_2 \geq \Delta_0$,

$$\text{Valor-p} = Pr(t \leq t_{teste}). \tag{5.40}$$

Sob $H_0 : \mu_1 - \mu_2 \leq \Delta_0$,

$$\text{Valor-p} = Pr(t \geq t_{teste}). \tag{5.41}$$

Exemplo 5.27. Adaptado de (Sheskin, 2011, p. 764).

```r
# dados
x1 <- c(9,2,1,4,6,4,7,8,5,1)
x2 <- c(8,2,3,2,3,0,4,5,4,0)
# validando suposições
shapiro.test(x1)  # S2, normalidade
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x1
## W = 0.9, p-value = 0.6
```

```r
shapiro.test(x2)  # S2, normalidade
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x2
## W = 0.9, p-value = 0.5
```

```r
g <- as.factor(rep(1:2, each = length(x1))) # grupos 1 e 2
car::leveneTest(c(x1,x2),g) # S3, homogeneidade de variâncias
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1    0.78   0.39
##       18
```

```r
# realizando o teste
D <- x1-x2
n <- length(D)
(mD <- mean(D))
```

```
## [1] 1.6
```

```r
(sD <- sd(D))
```

```
## [1] 1.78
```

```r
(tt <- mD/(sD/sqrt(n)))
```

```
## [1] 2.85
```

```r
gl <- n-1
2*(1-pt(tt, gl))
```

```
## [1] 0.0191
```

```r
# via t.test (facilita uma barbaridade!)
t.test(x1, x2, paired = T)
```

```
##
##  Paired t-test
##
## data:  x1 and x2
## t = 3, df = 9, p-value = 0.02
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.329 2.871
## sample estimates:
## mean of the differences
##                     1.6
```

Exercício 5.10. Faça o Exemplo 5.27 considerando $H_0 : \mu_1 - \mu_2 \geq 0$ e $H_0 : \mu_1 - \mu_2 \geq 0$.

TESTE 10 - Testes qui-quadrado para tabelas $l \times c$

Estes testes são extensões do teste qui-quadrado de aderência de uma amostra (TESTE 5).

Hipótese avaliada (geral)
Na(s) população(ões) representada(s) pela(s) amostra(s) em uma tabela de contingência, as frequências de células observadas são diferentes das frequências esperadas?

Hipóteses avaliadas (homogeneidade)
As l amostras são ou não homogêneas com relação à proporção de observações em cada uma das c categorias? (ou)
Se os dados forem homogêneos, a proporção de observações na j-ésima categoria será igual em todas as l populações?

Hipótese avaliada (independência)
As duas dimensões/variáveis são independentes uma da outra?

Suposições
S1. Os dados avaliados representam uma amostra aleatória composta por n observações independentes;
S2. Os dados de frequência são categóricos para $l \times c$ categorias mutuamente exclusivas;
S3. A frequência esperada de cada célula da tabela de contingência é de pelo menos 5;
S4. **(Homogeneidade)** As somas das linhas e colunas (somas marginais) **são** predeterminadas/fixas.
S4. **(Independência)** As somas das linhas e colunas (somas marginais) **não são** predeterminadas/fixas.

Testes relacionados No caso de $l > 2$ ou $c > 2$ pode-se considerar uma

análise *post hoc*[5] para o teste qui-quadrado de Pearson para dados de contagem, proposta por (Beasley and Schumacker, 1995), disponível no pacote `chisq.posthoc.test` (Ebbert, 2019).

Estatística do teste (sem correção de Yates)
Sob $H_0 : O_{ij} = E_{ij}$ para todas as células ou $H_0 : \pi_{ij} = (\pi_{i.})(\pi_{.j})$ para todas as $l \times c$ células,

$$\chi^2_{teste} = \sum_{i=1}^{l}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(gl), \tag{5.42}$$

onde

$$E_{ij} = \frac{(O_{i.})(O_{.j})}{n}. \tag{5.43}$$

k é o número de categorias e $gl = (l-1)(c-1)$ indica os *graus de liberdade* que definem a distribuição χ^2.

Estatística do teste (com correção de Yates)
Sob $H_0 : O_{ij} = E_{ij}$ para todas as células ou $H_0 : \pi_{ij} = (\pi_{i.})(\pi_{.j})$ para todas as $l \times c$ células,

$$\chi^2_{teste} = \sum_{i=1}^{l}\sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}} \sim \chi^2(gl), \tag{5.44}$$

Estatística do teste simplificada para tabelas 2×2 (sem correção de Yates)
Sob $H_0 : O_{ij} = E_{ij}$ para todas as células ou $H_0 : \pi_{ij} = (\pi_{i.})(\pi_{.j})$ para todas as $l \times c$ células,

$$\chi^2_{teste} = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}, \tag{5.45}$$

onde a, b, c e d são as quantidades conforme tabela a seguir.

	Coluna 1	Coluna 2	Total
Linha 1	a	b	**a+b**
Linha 2	c	d	**c+d**
Total	$a+c$	$b+d$	**n**

Estatística do teste simplificada para tabelas 2×2 (com correção de Yates)
Sob $H_0 : O_{ij} = E_{ij}$ para todas as células ou $H_0 : \pi_{ij} = (\pi_{i.})(\pi_{.j})$ para todas as

[5] A análise *post hoc* é aplicada quando uma hipótese de múltiplas igualdades é rejeitada, admitindo-se que 'há pelo menos dois grupos distintos'. Neste caso, fica a pergunta: quais grupos são distintos e em que grau? A terminologia é baseada na expressão latina que significa 'depois disto'.

$l \times c$ células,

$$\chi^2_{teste} = \frac{n(|ad - bc| - 0.5n)^2}{(a+b)(c+d)(a+c)(b+d)}. \tag{5.46}$$

Exemplo 5.28. (Adaptado de (Sheskin, 2011, 639), teste de homogeneidade) Um pesquisador realiza um estudo para avaliar o efeito do ruído no comportamento altruísta. Cada um dos 200 sujeitos que participam do experimento é atribuído aleatoriamente a uma de duas condições experimentais. Os indivíduos em ambas as condições realizam um teste de uma hora, que é ostensivamente uma medida de inteligência. Durante o teste, os 100 indivíduos do Grupo 1 são expostos a um ruído alto e contínuo, que, segundo eles, é devido a um gerador com defeito. Os 100 sujeitos do Grupo 2 não são expostos a nenhum ruído durante o teste. Após a conclusão desta etapa do experimento, cada sujeito, ao deixar a sala, é confrontado por um homem de meia-idade cujo braço está em uma tipóia e que trabalha no experimentado. O homem pergunta ao sujeito se estaria disposto a ajudá-lo a carregar um pacote pesado para o carro. O número de sujeitos em cada grupo que ajudam o homem é registrado. Trinta dos 100 sujeitos que foram expostos ao ruído optaram por ajudar o homem, enquanto sessenta dos 100 sujeitos que não foram expostos ao ruído optaram por ajudar o homem. Os dados indicam que o comportamento altruísta é influenciado pelo ruído?

	Ajudou	Não ajudou	Total
Barulho	30	70	**100**
Sem barulho	60	40	**100**
Total	**90**	**110**	**200**

```
dados <- matrix(c(30,60,70,40), nrow=2)
chisq.test(dados, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dados
## X-squared = 18, df = 1, p-value = 2e-05
```

```
chisq.test(dados, correct = T)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dados
## X-squared = 17, df = 1, p-value = 4e-05
```

Exemplo 5.29. (Adaptado de (Sheskin, 2011, 640), teste de independência)

Um pesquisador deseja testar se existe relação entre a dimensão da personalidade de introversão-extroversão e afiliação política. Duzentas pessoas são recrutadas para participar do estudo. Todos os sujeitos passam por um teste de personalidade com base no qual cada sujeito é classificado como introvertido ou extrovertido. Pede-se para cada sujeito indicar se ele ou ela é um democrata ou um republicano conforme tabela a seguir. Os dados indicam que existe uma relação significativa entre a afiliação política e se alguém é introvertido ou não extrovertido?

	Democrata	Republicano	Total
Introvertido	30	70	**100**
Extrovertido	60	40	**100**
Total	**90**	**110**	**200**

```
dados <- matrix(c(30,60,70,40), nrow=2)
chisq.test(dados, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dados
## X-squared = 18, df = 1, p-value = 2e-05
```

```
chisq.test(dados)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dados
## X-squared = 17, df = 1, p-value = 4e-05
```

Exercício 5.11. Faça os cálculos dos Exemplos 5.28 e 5.29 utilizando todas as variações de estatísticas de teste.

TESTE 11 - Teste exato de Fisher para tabelas 2×2

Este teste pode ser pensado como a versão exata (não assintótica) para o teste qui-quadrado de homogeneidade do TESTE 10 .

Hipótese avaliada (geral)
Na(s) população(ões) representada(s) pela(s) amostra(s) em uma tabela de contingência, as frequências de células observadas são diferentes das frequências esperadas?

Hipóteses avaliadas (homogeneidade)
As l amostras são ou não homogêneas com relação à proporção de observações

em cada uma das c categorias? (ou)
Se os dados forem homogêneos, a proporção de observações na j-ésima categoria será igual em todas as l populações?

Suposições
S1. Os dados avaliados representam uma amostra aleatória composta por n observações independentes;
S2. Os dados de frequência são categóricos para $l \times c$ categorias mutuamente exclusivas;
S3. As somas das linhas e colunas (somas marginais) **são** predeterminadas/fixas.

Valor-p para tabelas 2×2
Sob $H_0 : O_{ij} = E_{ij}$ para todas as células ou $H_0 : \pi_{ij} = (\pi_{i.})(\pi_{.j})$ para todas as $l \times c$ células,

$$P = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{n!a!b!c!d!}, \tag{5.47}$$

onde a, b, c e d são as quantidades conforme tabela a seguir.

	Coluna 1	Coluna 2	Total
Linha 1	a	b	**a+b**
Linha 2	c	d	**c+d**
Total	$a+c$	$b+d$	**n**

Exemplo 5.30. Pode-se resolver o Exemplo 5.28 através do teste exato de Fisher.

```
dados <- matrix(c(30,60,70,40), nrow=2)
fisher.test(dados)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  dados
## p-value = 3e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.152 0.534
## sample estimates:
## odds ratio
##      0.288
```

Exercício 5.12. Faça os cálculos do Exemplo 5.30.

5.3.3 Testes Paramétricos Multivariados

TESTE 12 - Análise de Variância (ANOVA) de um fator entre sujeitos

Hipótese avaliada
Em um conjunto de $k \geq 2$ grupos independentes, há pelo menos dois com médias distintas?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. A distribuição dos dados na população subjacente da qual cada uma das amostras é derivada é normal;
S3. (*Homogeneidade de variâncias*) A variância da população representada pelas k amostras são iguais entre si. ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$).

Testes relacionados
TESTE 19 - Teste de Kruskal-Wallis de um fator entre sujeitos

Estatística do teste
Sob $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$,

$$F_{teste} = \frac{MQ_{EG}}{MQ_{IG}} \sim \mathcal{F}(gl_{EG}, gl_{IG}), \tag{5.48}$$

onde MQ_{EG} é a média quadrática entre grupos dada por

$$MQ_{EG} = \frac{SQ_{EG}}{gl_{EG}}, \tag{5.49}$$

onde SQ_{EG} é a soma de quadrados entre grupos dada por

$$SQ_{EG} = \sum_{j=1}^{k} \left[\frac{(\sum x_j)^2}{n_j} \right] - \frac{(\sum x_T)^2}{n}, \tag{5.50}$$

MQ_{IG} é a média quadrática intra grupos dada por

$$MQ_{IG} = \frac{SQ_{IG}}{gl_{IG}}, \tag{5.51}$$

onde SQ_{IG} é a soma de quadrados intra grupos dada por

$$SQ_{IG} = \sum_{j=1}^{k} \left[\sum x_j^2 - \frac{(\sum x_j)^2}{n_j} \right], \tag{5.52}$$

$gl_{EG} = k-1$ são os graus de liberdade entre grupos, $gl_{IG} = n-k$ são os graus de liberdade intra grupos. Sabe-se também que onde SQ_{IG} é a soma de quadrados intra grupos dada por

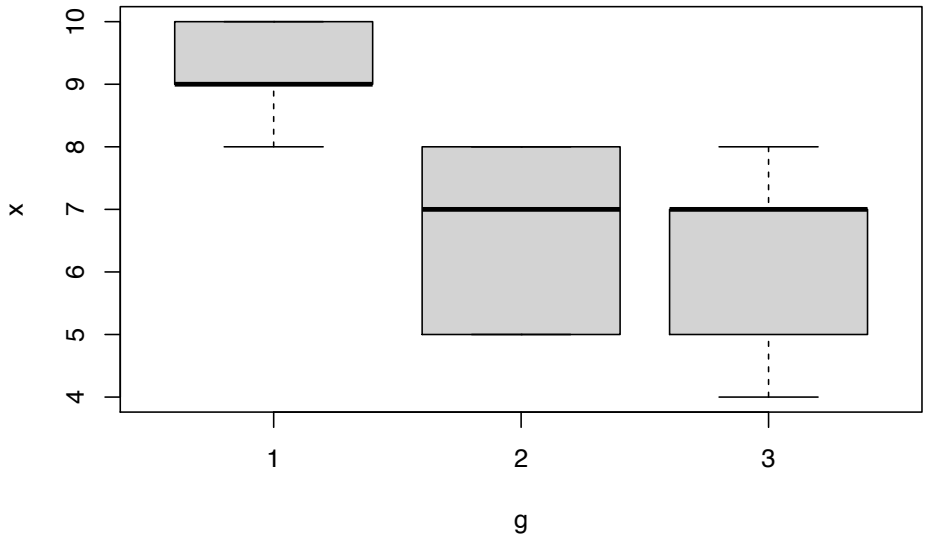
$$SQ_T = SQ_{EG} + SQ_{IG} = \sum x_T^2 - \frac{(\sum x_T)^2}{n}. \tag{5.53}$$

Exemplo 5.31. (Adaptado de (Sheskin, 2011, 886)) Um psicólogo realiza um estudo para determinar se o ruído pode ou não inibir o aprendizado. Cada um de 15 sujeitos é atribuído aleatoriamente a um dos três grupos. Cada sujeito tem 20 minutos para memorizar uma lista de 10 sílabas sem sentido, que ela diz que será testada no dia seguinte. Os cinco sujeitos atribuídos ao Grupo 1, a condição sem ruído, estudam a lista de sílabas sem sentido enquanto estão em uma sala silenciosa. Os cinco sujeitos designados para o Grupo 2, a condição de ruído moderado, estudam a lista de sílabas sem sentido enquanto ouvem música clássica. Os cinco sujeitos designados para o Grupo 3, a condição de ruído extremo, estudam a lista de sílabas sem sentido enquanto ouvem música rock. O número de sílabas sem sentido lembradas corretamente pelos 15 sujeitos segue: Grupo 1: 8,10,9,10,9; Grupo 2: 7,8,5,8,5; Grupo 3: 4,8,7,5,7. Os dados indicam que o ruído influenciou o desempenho dos sujeitos?

```
# dados
x <- c(8,10,9,10,9, 7,8,5,8,5, 4,8,7,5,7)
g <- as.factor(rep(1:3, each = 5))
(k <- length(unique(g))) # número de grupos
```

```
## [1] 3
```

```
boxplot(x ~ g)
```




```
# validando suposições, veja ?by
by(x,g,shapiro.test)  # S2, normalidade
```

```
## g: 1
##
##  Shapiro-Wilk normality test
##
```

```
## data:  dd[x, ]
## W = 0.9, p-value = 0.3
##
## ------------------------------------------------------------------------------------
## g: 2
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.8, p-value = 0.09
##
## ------------------------------------------------------------------------------------
## g: 3
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9, p-value = 0.5
```

```
car::leveneTest(x,g)  # S3, homogeneidade de variâncias
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2    0.67   0.53
##       12
```

```
# cálculos
(soma <- by(x,g,sum))     # soma por grupo
```

```
## g: 1
## [1] 46
## ------------------------------------------------------------------------------------
## g: 2
## [1] 33
## ------------------------------------------------------------------------------------
## g: 3
## [1] 31
```

```
(media <- by(x,g,mean))   # média por grupo
```

```
## g: 1
## [1] 9.2
## ------------------------------------------------------------------------------------
## g: 2
## [1] 6.6
## ------------------------------------------------------------------------------------
## g: 3
## [1] 6.2
```

```
(soma2 <- by(x^2,g,sum))  # soma ao quadrado por grupo
```

```
## g: 1
## [1] 426
## ------------------------------------------------------------------------------------------------------------
## g: 2
## [1] 227
## ------------------------------------------------------------------------------------------------------------
## g: 3
## [1] 203
```

```
(n <- by(x,g,length))     # tamanho da amostra por grupo
```

```
## g: 1
## [1] 5
## ------------------------------------------------------------------------------------------------------------
## g: 2
## [1] 5
## ------------------------------------------------------------------------------------------------------------
## g: 3
## [1] 5
```

```
(somaT <- sum(soma))      # soma total
```

```
## [1] 110
```

```
(soma2T <- sum(soma2))    # soma ao quadrado total
```

```
## [1] 856
```

```
(nT <- sum(n))            # tamanho total da amostra
```

```
## [1] 15
```

```
(sqt <- soma2T - somaT^2/nT) # SQ_T pela Eq. (3.50)
```

```
## [1] 49.3
```

```
(sqeg <- sum(soma^2/n) - somaT^2/nT) # SQ_{EG} pela Eq. (3.47)
```

```
## [1] 26.5
```

```
(sqig <- sum(soma2 - soma^2/n)) # SQ_{IG} pela Eq. (3.49)
```

```
## [1] 22.8
```

```
sqt - sqeg # SQ_{IG} pela Eq. (3.50)
```

```
## [1] 22.8
```

```
(gleg <- k-1) # gl_{EG}
```

```
## [1] 2
```

```
(glig <- nT-k) # gl_{IG}
```

```
## [1] 12
```

```
(mqeg <- sqeg/gleg) # MQ_{EG} pela Eq. (3.46)
```

```
## [1] 13.3
```

```
(mqig <- sqig/glig) # MQ_{IG} pela Eq. (3.48)
```

```
## [1] 1.9
```

```
(Ft <- mqeg/mqig) # estatística do teste pela Eq. (3.45)
```

```
## [1] 6.98
```

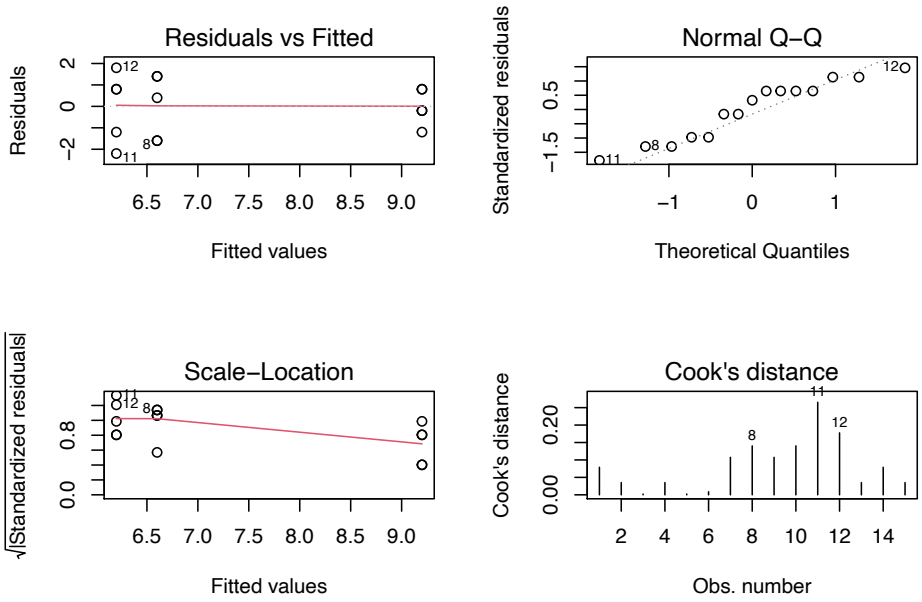
```
2*df(Ft,gleg,glig)
```

```
## [1] 0.00901
```

```
# pela função aov
summary(anova1 <- aov(x ~ g))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## g            2   26.5    13.3    6.98 0.0097 **
## Residuals   12   22.8     1.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
plot(anova1, which=1:4)
```




```
# pela função anova
anova(lm(x ~ g))
```

```
## Analysis of Variance Table
##
## Response: x
##           Df Sum Sq Mean Sq F value Pr(>F)
## g          2   26.5    13.3    6.98 0.0097 **
## Residuals 12   22.8     1.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# pela função car::Anova
car::Anova(lm(x ~ g))
```

```
## Anova Table (Type II tests)
##
## Response: x
##           Sum Sq Df F value Pr(>F)
## g           26.5  2    6.98 0.0097 **
## Residuals   22.8 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# pela função lm
summary(lm(x ~ g))
```

```
##
```

```
## Call:
## lm(formula = x ~ g)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.4    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.200      0.616   14.92  4.1e-09 ***
## g2            -2.600      0.872   -2.98   0.0114 *
## g3            -3.000      0.872   -3.44   0.0049 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.38 on 12 degrees of freedom
## Multiple R-squared:  0.538,  Adjusted R-squared:  0.461
## F-statistic: 6.98 on 2 and 12 DF,  p-value: 0.00974
```

```
# Post-hoc
stats::TukeyHSD(anova1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x ~ g)
##
## $g
##     diff   lwr    upr p adj
## 2-1 -2.6 -4.93 -0.274 0.029
## 3-1 -3.0 -5.33 -0.674 0.013
## 3-2 -0.4 -2.73  1.926 0.891
```

```
rstatix::tukey_hsd(anova1)
```

```
## # A tibble: 3 x 9
##   term  group1 group2 null.value estimate conf.low conf.high  p.adj p.adj.signif
## * <chr> <chr>  <chr>       <dbl>    <dbl>    <dbl>     <dbl>  <dbl> <chr>
## 1 g     1      2               0    -2.60    -4.93    -0.274 0.0286 *
## 2 g     1      3               0    -3       -5.33    -0.674 0.0126 *
## 3 g     2      3               0    -0.4     -2.73     1.93  0.891  ns
```

Exercício 5.13. (Adaptado de (DeGroot and Schervish, 2012, p. 754)) Moore e McCabe (1999) descrevem dados coletados em *Consumer Reports* (junho de 1986, pp. 364-67). Os dados incluem (entre outras coisas) calorias conteúdo de 63 marcas de salsichas de cachorros-quentes conforme tabela a seguir. A salsichas vêm em quatro variedades: carne bovina, carne (?), aves e especialidades. É

interessante saber se, e em que medida, as diferentes variedades diferem em seus conteúdos calóricos. Realize o procedimento de análise de variância e post hoc, indicando se há diferença significativa entre os grupos.

Carne bovina 186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132
Carne 173, 191, 182, 190, 172, 147, 146, 139, 175, 136, 179, 153,107, 195, 135, 140, 138
Aves 129, 132, 102, 106, 94, 102, 87, 99, 107, 113, 135, 142, 86, 143, 152, 146, 144
Especialidades 155, 170, 114, 191, 162, 146, 140, 187, 180

```
# Dica
g <- as.factor(rep(1:4, times=c(20,17,17,9)))
```

5.3.4 Testes Não Paramétricos Univariados

TESTE 13 - Teste dos postos sinalizados de Wilcoxon para uma amostra

Hipótese avaliada
Uma amostra de n sujeitos (ou objetos) vem de uma população em que a mediana θ é igual a um valor especificado?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. As pontuações originais obtidas para cada um dos sujeitos/objetos são quantitativas;
S3. A distribuição da população subjacente é simétrica.

Testes relacionados
TESTE 1 - Teste z para média de uma amostra
TESTE 2 - Teste t para média de uma amostra

Para testar a simetria foi considerada a função `symmetry_test` do pacote `symmetry`. Segundo (Milošević and Obradović, 2018, p. 4), entre os testes originalmente destinados a testar simetria em torno de uma média desconhecida, o mais famoso é o teste clássico $\sqrt{b_1}$, baseado no coeficiente de assimetria da amostra, com estatística de teste $\sqrt{b_1} = m_3/s^3$ onde m_3 é o terceiro momento central da amostra conforme numerador da Eq. (2.25) e s é o desvio padrão amostral conforme Eq. (2.22).

```
# dados
set.seed(456); z <- rnorm(100) # N(0,1)
# verificando suposição S3 (simetria)
library(symmetry)
set.seed(111); symmetry_test(z, 'B1')  # √b1
```

```
##
```

```
##  Symmetry test
##  Null hypothesis: Data is symmetric
##
## data:  z
## B1 = -2e-04, B = 1000, p-value = 1
## sample estimates:
##    mu
## 0.121
```

```
# teste de Wilcoxon
wilcox.test(z, mu = 0) # mediana 0
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  z
## V = 2851, p-value = 0.3
## alternative hypothesis: true location is not equal to 0
```

```
wilcox.test(z, mu = 1) # mediana 1
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  z
## V = 563, p-value = 2e-11
## alternative hypothesis: true location is not equal to 1
```

TESTE 14 - Teste de aderência de Shapiro-Wilk para uma amostra

Hipótese avaliada
Os dados têm distribuição normal?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. A escala de mensuração é quantitativa.

```
set.seed(8765); z <- rnorm(100, mean = 5, sd = 3)
shapiro.test(z)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  z
## W = 1, p-value = 0.3
```

```
set.seed(7654); u <- runif(100, min = 4, max = 6)
shapiro.test(u)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  u
## W = 1, p-value = 0.002
```

5.3.5 Testes Não Paramétricos Bivariados

TESTE 15 - Teste de Kolmogorov-Smirnov para duas amostras independentes

Hipótese avaliada
Duas amostras independentes representam duas populações distintas?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. A escala de mensuração é pelo menos ordinal.

```
# veja ?ks.test
set.seed(99); x <- rnorm(50)
set.seed(88); y <- runif(30)
ks.test(x, y)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.5, p-value = 4e-05
## alternative hypothesis: two-sided
```

```
set.seed(77); z <- runif(30)
ks.test(x, z)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and z
## D = 0.5, p-value = 1e-05
## alternative hypothesis: two-sided
```

```
ks.test(y, z)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  y and z
## D = 0.2, p-value = 0.8
## alternative hypothesis: two-sided
```

TESTE 16 - Teste dos postos de Mann-Whitney para amostras independentes

Hipótese avaliada
Duas amostras independentes representam duas populações com medianas diferentes?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. As duas amostras são independentes entre si;
S3. Os dados são ordinais ou quantitativos;
S4. As distribuições de onde as amostras foram retiradas possuem mesma forma.

Testes relacionados
TESTE 7 - Teste z para médias de duas amostras independentes
TESTE 8 - Teste t para médias de duas amostras independentes

Exemplo 5.32. (Adaptado de (Sheskin, 2011, 532)) Para avaliar a eficácia de um novo medicamento antidepressivo, dez pacientes com depressão clínica são aleatoriamente designados para um dos dois grupos. Cinco pacientes são atribuídos ao Grupo 1, onde é administrado o antidepressivo por um período de seis meses. Os outros cinco pacientes são atribuídos ao Grupo 2, que recebe um placebo durante o mesmo período de seis meses. Suponha que, antes de introduzir os tratamentos experimentais, o experimentador confirmou que o nível de depressão nos dois grupos era igual. Após seis meses, todos os dez sujeitos são avaliados por um psiquiatra (que é cego em relação à condição experimental do sujeito) quanto ao nível de depressão. As classificações de depressão do psiquiatra para os cinco sujeitos em cada grupo seguem (quanto mais alta a classificação, mais deprimido é o sujeito): Grupo 1: 11, 1,0, 2, 0; Grupo 2: 11, 11, 5, 8, 4. Os dados indicam que o antidepressivo é eficaz?

$$\begin{cases} H_0 : \theta_1 \geq \theta_2 \equiv \theta_1 - \theta_2 \geq 0 \text{ (tratamento igual ou menos eficaz que o placebo)} \\ H_1 : \theta_1 < \theta_2 \equiv \theta_1 - \theta_2 < 0 \text{ (tratamento mais eficaz que o placebo)} \end{cases}$$

```
x <- c(11,1,0,2,0)
y <- c(11,11,5,8,4)
median(x); median(y)
```

```
## [1] 1
```

```
## [1] 8
```

```
# avaliando S4
ks.test(x,y)
```

```
## Warning in ks.test(x, y): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.8, p-value = 0.08
## alternative hypothesis: two-sided
```

```
wilcox.test(x, y, alternative  = 'less')
```

```
## Warning in wilcox.test.default(x, y, alternative = "less"): cannot compute exact p-value with
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 4, p-value = 0.04
## alternative hypothesis: true location shift is less than 0
```

TESTE 17 - Teste dos postos de Wilcoxon para amostras dependentes/pareadas

Hipótese avaliada
Duas amostras dependentes representam duas populações distintas?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. As pontuações originais obtidas para cada um dos sujeitos/objetos são quantitativas;
S3. A distribuição dos escores de diferença nas populações representadas pelas duas amostras é simétrica em relação à mediana da população de escores de diferença.

Testes relacionados
TESTE 9 - Teste t para médias de duas amostras dependentes/pareadas.

Assim como no *teste dos postos sinalizados de Wilcoxon para uma amostra* (TESTE 14), para testar a simetria foi considerada a estatística $\sqrt{b_1}$ da função `symmetry::symmetry_test`.

```
# dados
x <- c(1.83,  0.50,  1.62,  2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)
median(x-y)
```

```
## [1] 0.49
```

```
# verificando suposição S3 (simetria da diferença em relação à mediana)
set.seed(111); symmetry::symmetry_test(x-y, 'B1', mu = median(x-y))  # √b1
```

```
##
```

```
##  Symmetry test
##  Null hypothesis: Data is symmetric around 0.49
##
## data:  x - y
## B1 = -0.07, B = 1000, p-value = 0.9
```

```
# teste de Wilcoxon para amostras pareadas
wilcox.test(x, y, paired = TRUE, alternative = 'greater')
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  x and y
## V = 40, p-value = 0.02
## alternative hypothesis: true location shift is greater than 0
```

TESTE 18 - Teste de McNemar para amostras dependentes/pareadas

Hipótese avaliada
Duas amostras dependentes representam duas populações distintas?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. As variáveis são binárias e categóricas (ordinais ou nominais);
S3. Cada um dos n sujeitos (ou n pares de sujeitos combinados) contribui com duas pontuações na variável dependente.

Estatística do teste
Sob $H_0 : \pi_{12} = \pi_{21}$,

$$\chi^2_{teste} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2(1) \tag{5.54}$$

onde n_{12} indica o número de elementos da linha 1, coluna 2 da tabela e n_{21} indica o número de elementos da linha 2, coluna e $\chi^2(1)$ indica a distribuição qui-quadrado com 1 grau de liberdade.

Estatística do teste com correção de continuidade
Sob $H_0 : \pi_{12} = \pi_{21}$,

$$\chi^2_{teste} = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}} \sim \chi^2(1) \tag{5.55}$$

Valor-p
Sob $H_0 : \pi_{12} = \pi_{21}$,

$$\text{Valor-p} = Pr(\chi^2 \geq \chi^2_{teste}). \tag{5.56}$$

Exemplo 5.33. (Exemplo da documentação de `mcnemar.test`) A aprovação do desempenho do Presidente no cargo foi realizado em duas pesquisas, com um mês de intervalo, para uma amostra aleatória de 1.600 americanos em idade

de votar. Os dados indicam uma mudança de percepção em relação às duas pesquisas?

```
# dados
dat <- matrix(c(794, 86, 150, 570), nrow = 2,
              dimnames = list('1ª pesquisa' = c('Aprova', 'Desaprova'),
                              ' 2ª pesquisa' = c('Aprova', 'Desaprova')))
dat
```

```
##             2ª pesquisa
## 1ª pesquisa Aprova Desaprova
##   Aprova       794       150
##   Desaprova     86       570
```

```
# usando a estatística do teste sem correção
(qui <- (dat[1,2]-dat[2,1])^2/(dat[1,2]+dat[2,1]))
```

```
## [1] 17.4
```

```
# valor-p
1-pchisq(qui,1)
```

```
## [1] 3.1e-05
```

```
# teste via mcnemar.test
mcnemar.test(dat, correct = F)
```

```
##
##  McNemar's Chi-squared test
##
## data:  dat
## McNemar's chi-squared = 17, df = 1, p-value = 3e-05
```

```
# usando a estatística do teste com correção
(qui.c <- (abs(dat[1,2]-dat[2,1])-1)^2/(dat[1,2]+dat[2,1]))
```

```
## [1] 16.8
```

```
# valor-p
1-pchisq(qui.c,1)
```

```
## [1] 4.11e-05
```

```
# teste via mcnemar.test
mcnemar.test(dat, correct = T)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  dat
## McNemar's chi-squared = 17, df = 1, p-value = 4e-05
```

Exercício 5.14. Um pesquisador conduz um estudo para investigar se uma série semanal de televisão altamente crítica quanto ao uso de animais em pesquisas médicas influencia a opinião pública. Cem sujeitos selecionados aleatoriamente são avaliados por um teste para determinar sua atitude em relação ao uso de animais em pesquisas médicas. Com base em suas respostas, os sujeitos são então categorizados como pró-pesquisa com animais ou anti-pesquisa com animais. Após o pré-teste, todos os sujeitos são orientados a assistir à série de televisão (com duração de dois meses). Na conclusão da série, a atitude de cada sujeito em relação à pesquisa animal é reavaliada. Os resultados do estudo estão resumidos na tabela a seguir. Os dados indicam que uma mudança de atitude em relação à pesquisa com animais ocorreu depois que os participantes assistiram à série de televisão?
Resolva utilizando as estatísticas com e sem correção de continuidade, realizando os cálculos e também aplicando a função `mcnemar.test`.

	Pós-teste		
Pré-teste	**Anti**	**Pró**	**Total**
Anti	10	13	**23**
Pró	41	36	**77**
Total	**51**	**49**	**100**

5.3.6 Testes Não Paramétricos Multivariados

TESTE 19 - Teste de Kruskal-Wallis de um fator entre sujeitos

Hipótese avaliada
Em um conjunto de $k \geq 2$ grupos independentes, há pelo menos dois com medianas distintas?

Suposições
S1. Cada amostra foi selecionada aleatoriamente da população que representa;
S2. As k amostras são independentes umas das outras;
S3. A variável dependente (que é subsequentemente classificada) é uma variável aleatória contínua.
S4. As distribuições subjacentes das quais as amostras são derivadas possuem a mesma forma, não obrigatoriamente normais.

Testes relacionados
TESTE 12 - Análise de Variância (ANOVA) de um fator entre sujeitos

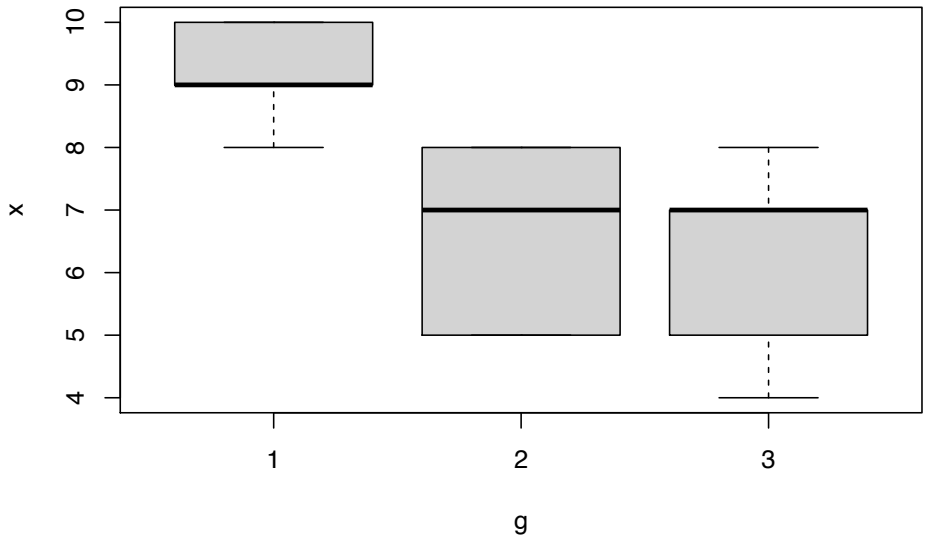
Exemplo 5.34. Pode-se resolver o Exemplo 5.31 através do teste de Kruskal-Wallis.

```
# dados
x <- c(8,10,9,10,9, 7,8,5,8,5, 4,8,7,5,7)
```

```r
g <- as.factor(rep(1:3, each = 5))
(k <- length(unique(g))) # número de grupos
```

```
## [1] 3
```

```r
boxplot(x ~ g)
```




```r
# validando suposição S4 via TESTE 19 - teste K-S para duas amostras independentes
ks.test(x[1:5], x[6:10])
```

```
## Warning in ks.test(x[1:5], x[6:10]): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x[1:5] and x[6:10]
## D = 0.8, p-value = 0.08
## alternative hypothesis: two-sided
```

```r
ks.test(x[1:5], x[11:15])
```

```
## Warning in ks.test(x[1:5], x[11:15]): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x[1:5] and x[11:15]
## D = 0.8, p-value = 0.08
## alternative hypothesis: two-sided
```

```
ks.test(x[6:10], x[11:15])
```

```
## Warning in ks.test(x[6:10], x[11:15]): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x[6:10] and x[11:15]
## D = 0.2, p-value = 1
## alternative hypothesis: two-sided
```

```
# aplicando o teste de Kruskal-Wallis
kruskal.test(x,g)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  x and g
## Kruskal-Wallis chi-squared = 9, df = 2, p-value = 0.01
```

```
# comparando com a ANOVA
summary(aov(x ~ g))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## g            2   26.5    13.3    6.98 0.0097 **
## Residuals   12   22.8     1.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercício 5.15. Refaça o exercício 5.13 aplicando o teste de Kruskal-Wallis.

5.4 Exercícios

1. O diâmetro nominal de uma amostra de 20 peças, cujos resultados estão em cm foram:

41	50	52	49	49	54	50	47	52	49
50	52	50	47	49	51	46	50	49	50

(a) Suponha inicialmente que o diâmetro populacional possui variância de $\sigma^2 = 2$ cm^2. Teste a hipótese de que a média seja diferente de 50cm com $\alpha = 0.05$.
(b) Faça o mesmo teste para a média, mas agora desconhecendo a variância, novamente com $\alpha = 0.05$.

2. Um processo deveria produzir mesas com 0.85m de altura. O engenheiro desconfia que as mesas que estão sendo produzidas são menores que o

especificado. Uma amostra de 8 mesas foi coletada e indicou média 0.847m. Sabendo que o desvio padrão é $\sigma = 0.01$m, obtenha o valor-p e teste a hipótese do engenheiro usando um nível de significância de 3%.

3. As condições de mortalidade de uma região são tais que a proporção de nascidos que sobrevivem até 60 anos é de 0.6. Testar essa hipótese ao nível de 5% se em 1000 nascimentos amostrados aleatoriamente, verificou-se 530 sobreviventes até 60 anos.

4. A experiência tem comprovado que mais de 40% dos estudantes são reprovados em uma prova de certa matéria. Se 45 de 90 estudantes amostrados fossem reprovados, o que se pode concluir a respeito desta afirmação. Obtenha o valor-p e teste esta hipótese ao nível de significância de 4%.

5. Funcionários de uma grande firma de contabilidade alegam que seu salário médio anual é menor que o de seu concorrente que é de R$ 45,000, sabe-se também que o desvio padrão de ambas as empresas são iguais a R$ 5,200. Uma amostra de 30 contadores da empresa gera um salário médio de R$ 43,500. Teste a alegação dos empregados ao nível de significância de 1%.

6. Certo fabricante de parafusos anuncia que 90% do seu produto não apresenta qualquer tipo de defeito. Um comprador acredita que a percentagem de parafusos perfeitos é diferente da anunciada pelo fabricante. Para verificar tal hipótese, examinou 400 parafusos, verificando que 344 eram perfeitos. Com $\alpha = 2\%$, realize o teste correspondente.

7. Certa organização médica afirma que um novo medicamento é de qualidade superior ao até então existente, que é 80% eficaz na cura de determinada doença. Examinada uma amostra de 300 pessoas que sofriam da doença, constatou-se que 249 ficaram curadas com o novo medicamento. Obtenha o valor-p e teste a afirmação da organização ao nível de significância de 5%.

8. Como responsável pelas compras em um mercado, suponha que você tome uma amostra aleatória de 32 latas de um certo produto. O peso líquido encontrado foi de 15.95g e o desvio padrão de 0.15g. Foi afirmado que o peso médio desse produto é 16.0g. Pode-se considerar essa afirmação verdadeira ao nível de significância de 5%? Usando a tabela t, o valor-p estaria entre quais valores? E usando o software, qual o valor mais preciso?

9. Uma certa agência bancária afirma que o tempo médio de espera na fila é de 15 minutos. Entretanto, os clientes estão revoltados com a demora no atendimento e dizem que a afirmação não é verdadeira, ou seja, o tempo de espera é superior a 15 min. Para poder argumentar contra o banco, os clientes realizam uma amostra com 200 pessoas, anotando o tempo até o atendimento. O resultado foi um tempo médio de espera de 19 min e variância de 49 min^2. Ao nível de significância de 5%, quem tem razão?

10. A empresa XYZ possui o seguinte critério para aceitar ou rejeitar um lote de matéria prima (aço). Se o nível de pureza do aço for superior a 90.0 o

lote é aceito, caso contrário é rejeitado. Um lote chega à empresa e cinco amostras são analisadas. Os níveis de pureza encontrados estão apresentados a seguir. O lote deve ser aceito, utilizando nível de significância de 1%? Formule as hipóteses da forma que achar mais adequada.

89.8	89.0	92.2	92.5	93.0

11. Um estudo do tempo médio de adaptação dos funcionários está sendo realizado num grande complexo industrial. Supõe-se que os homens tenham um tempo de adaptação menor do que as mulheres. Sabendo que numa amostra de 31 homens o tempo médio de adaptação foi de 3.2 anos e o desvio padrão de 1.3 ano. Numa amostra de 41 mulheres o tempo médio de adaptação foi de 3.7 anos e desvio de 0.8 ano.

(a) É possível obter o p-valor exato?
(b) Teste essa hipótese com $\alpha = 0.05$.

12. Um estudo está sendo realiza com com 121 crianças de escolas da rede pública e 121 crianças da rede particular. O estudo consiste da realização de um jogo onde é anotado o tempo de realização até a obtenção da resposta correta. Nas crianças da rede pública o tempo médio foi de 12 minutos, com um desvio padrão de 8 minutos. Na rede particular o tempo médio foi de 13 minutos, com um desvio de 5 minutos. Teste ao nível de significância de 5% se há diferença entre os tempos de realização do jogo entre alunos da rede pública e particular.

13. A empresa Chips está oferecendo a suas melhores equipes de vendas, prêmios em forma de viagens. A empresa quer saber se realmente os prêmios estão incentivando seus vendedores. Segue amostra de 8 equipes de vendas com o volume de vendas em ABR e MAI (considere que os meses de abril e maio são historicamente semelhantes).

Equipe	A	B	C	D	E	F	G	H
Abril (antes da promoção)	15.7	18.8	19.7	16.8	59.1	2.7	45.1	3.1
Maio (após a promoção)	17.0	18.7	21.5	17.6	65.2	2.5	47.2	3.7

Verifique se a campanha publicitária pode ser considerada eficiente ao nível de 1%. Qual o valor-p?

14. Por determinação do Governo Federal, as indústrias refinadoras de Sal devem misturar 1mg (0.001g) de iodo para cada grama de sal. Suspeita-se que a marca SALMOURA DOS PAMPAS não esteja cumprindo a especificação. Na amostra abaixo foram observados os resultados (em mg). Qual sua decisão com $\alpha = 0.05$?

1.2	1.1	1.01	0.9	0.8

15. As lojas GASTÃO e PRAQUETÁ pertencem a um mesmo grupo. Suspeita-se que o nível de satisfação médio dos clientes seja diferente de uma loja para outra. Os dados a seguir são baseados numa pesquisa feita por telefone que utilizou uma escala de 0 (pior avaliação) até 100 (melhor avaliação). Compare o nível de satisfação das lojas utilizando um nível de significância de 5%.

GASTÃO	PRAQUETÁ
$n = 180$	$n = 160$
$\bar{x}_G = 85.0$	$\bar{x}_P = 77.0$
$s_G = 18.0$	$s_P = 15.0$

16. A frequência crítica de oscilação (fco) é a frequência mais alta (em ciclos/s) na qual uma pessoa pode detectar a oscilação em uma fonte de luz oscilante. Em frequências acima de fco, a fonte de luz parece ser contínua, embora na verdade esteja oscilante. Uma investigação realizada para verificar se a fco média real depende da cor da íris, gerou os dados abaixo. Investigue as diferenças entre as cores em relação à fco média, com $\alpha = 5\%$.

Cor	Dados							
Marrom	26.8	27.9	23.7	25.0	26.3	24.8	24.5	25.7
Verde	26.4	24.2	28.0	26.9	29.1			
Azul	25.7	27.2	29.9	28.5	29.4	28.3		

Chapter 6

Inferência Bayesiana

O paradigma da *inferência bayesiana* tem suas origens no artigo póstumo de (Bayes, 1763), comunicado por seu amigo Richard Price. As derivações das ideias de Bayes são extensas e profundas matemática e filosoficamente, discutidas por grandes nomes da Ciência em incontáveis livros, artigos e compilações ao longo destes mais de 250 anos. Desta forma entende-se que a melhor abordagem para este material é indicar o estado-da-arte da aplicação bayesiana considerando referências consagradas disponíveis online ou na bilbioteca da PUCRS.

Exercício 6.1. Assita aos vídeos The Bayesian Trap e Bayes theorem. Lembre que você pode ativar a legenda (botão CC) e alterar a língua nas configurações (ícone de engrenagem > legendas).

Um dos principais motivos dos avanços recentes na pesquisa em estatística bayesiana é a crescente facilidade no acesso a recursos computacionais, tanto de hardware quanto de software. Na linguagem R existem muitas bibliotecas para aplicação bayesiana. O CRAN Task View[1] de inferência bayesiana fornece um compêndio atualizado dos pacotes relacionados ao assunto.

[1]Segundo a documentação oficial do R, os *CRAN Task Views* ('Visualizadores de Tarefa da Rede Abrangente de Arquivos R', em tradução livre) têm como objetivo fornecer alguma orientação sobre quais pacotes no CRAN são relevantes para tarefas relacionadas a um determinado tópico. Eles fornecem uma breve visão geral dos pacotes incluídos têm como objetivo ter um foco nítido para que seja suficientemente claro quais pacotes devem ser incluídos (ou excluídos) - e não têm a intenção de endossar os "melhores" pacotes para uma determinada tarefa.

6.1 Princípios de verossimilhança, suficiência e condicionalidade

- Seção 1.6 de (Paulino et al., 2003)
- Seções 3.3 e 3.4 de (Press, 2003) (Princípio da verossimilhança)
- Os fundamentos são discutidos por (Birnbaum, 1962), (Savage et al., 1962) e (Wechsler et al., 2008)

Informalmente, o *princípio da verossimilhança* admite que se dois decisores possuem o mesmo grau de conhecimento e a mesma informação sobre θ, ambos devem decidir exatamente da mesma forma a respeito de θ. (Berger, 1985, p.28)[2] define da seguinte forma:

> **Princípio da verossimilhança** Ao fazer inferências ou decisões sobre θ após observar x, todas as informações experimentais relevantes estão contidas na função de verossimilhança para o x observado. Além disso, duas funções de verossimilhança contêm as mesmas informações sobre θ se forem proporcionais entre si (como funções de θ).

Exemplo 6.1. (Princípio da Verossimilhança 1, adaptado de (Paulino et al., 2003)) Considere uma sucessão de lançamentos de uma moeda, independentes e condicionados por θ, a probabilidade de sair 'cara'. Suponha que seja obtido o resultado

$$x = \{H, T, H, H, T, T, H, T, T, T\},$$

onde H: 'cara' e T: 'coroa'[3]. Este resultado poderia ser obtido de diversos processos experimentais ou regras de parada, como
- realizar 10 lançamentos, fixados a priori
- lançar a moeda até aparecerem 6 'coroas'
- lançar a moeda até aparecerem 3 'coroas' consecutivas
- lançar a moeda até o jogador ficar saturado, tendo a saturação ocorrido no 10º lançamento

Em qualquer caso a (função de) verossimilhança é proporcional a $\theta^4 (1-\theta)^6$, i.e., a amostra informa quatro sucessos (caras) e seis fracassos (coroas). Assim, adotando-se o princípio da verossimilhança, toda a informação que x pode fornecer sobre θ encontra-se nesta expressão. Saber qual dos quatro processos experimentais foi utilizado (cada um com um espaço amostral diferente) ou saber qual foi a regra de parada adotada nada tem a acrescentar. Note que

[2] "*The Likelihood Principle. In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ).*" (Berger, 1985, p.28)

[3] Do Inglês *Head* (cara) e *Tail* (coroa).

a possibilidade de o experimentador parar por seu arbítrio ao considerar o resultado x satisfatório, em nada altera a opinião sobre θ.

Exemplo 6.2. (Princípio da Verossimilhança 2, adaptado de (Lindley and Phillips, 1976) por (Paulino et al., 2003)) Suponha que deseja-se testar a hipótese $H_0 : \theta \leq 1/2$ contra $H_1 : \theta > 1/2$. São contemplados dois processos experimentais:

- E_1: lançar a moeda $n = 12$ vezes;
- E_2: lançar a moeda até que apareçam $k = 3$ 'caras'

Admita que o resultado observado nas duas experiências foi $x = 9$ 'coroas' (portanto 3 'caras'), que é uma particular realização da variável aleatória X, que designa o número total de 'coroas' dos experimentos E_1 e E_2. Para um clássico o nível crítico (ou valor-p, a probabilidade de obter $X \geq 9$) da hipótese $H_0 : \theta = 1/2$ difere nos dois casos.

No caso E_1, X tem distribuição binomial – $X \sim \mathcal{B}(12, \theta)$ – cujo nível crítico é

$$Pr\left(X \geq 9 \middle| \theta = \frac{1}{2}\right) = \binom{12}{9}\left(\frac{1}{2}\right)^{12} + \binom{12}{10}\left(\frac{1}{2}\right)^{12} + \binom{12}{11}\left(\frac{1}{2}\right)^{12} + \binom{12}{12}\left(\frac{1}{2}\right)^{12} \approx 0.0730.$$

No caso E_2, X tem distribuição binomial negativa – $X \sim \mathcal{BN}(3, 1-\theta)$ – que tem nível crítico

$$Pr\left(X \geq 9 \middle| \theta = \frac{1}{2}\right) = \binom{11}{9}\left(\frac{1}{2}\right)^{12} + \binom{12}{10}\left(\frac{1}{2}\right)^{13} + \binom{13}{11}\left(\frac{1}{2}\right)^{14} + \cdots \approx 0.0327.$$

Logo, se for adotado um limiar de significância de 5%, H_0 é rejeitada no caso E_2 e não rejeitada em E_1. Assumindo o princípio da verossimilhança, as conclusões devem ser idênticas nos dois casos. Em ambos a (função de) verossimilhança é proporcional a $\theta^9 (1-\theta)^3$. De fato, as verossimilhanças em E_1 e E_2 são

$$L_1(\theta | \, x = 9) = \binom{12}{9}\theta^9 (1-\theta)^3 = 220\, \theta^9 (1-\theta)^3 \propto \theta^9 (1-\theta)^3$$

$$L_2(\theta | \, x = 9) = \binom{11}{9}\theta^9 (1-\theta)^3 = 55\, \theta^9 (1-\theta)^3 \propto \theta^9 (1-\theta)^3$$

6.2 Distribuição a priori

- Fundamentos abordados no Capítulo 2 de (Paulino et al., 2003) e no Capítulo 5 de (Press, 2003)
- (Morris et al., 2014) apresentam uma ferramenta baseada na web para obter distribuições de probabilidade de especialistas

6.3 Estimação Pontual

- Seções 8.2 e 8.3 de (Press, 2003)
- Seção 3.2 de (Paulino et al., 2003)

6.4 (Estimação por) Intervalo/Regiões de Credibilidade

- Seção 8.4 de (Press, 2003)
- Seção 3.3 de (Paulino et al., 2003)

6.5 (Estimação por) Teste de Hipóteses

- Capítulo 9 de (Press, 2003)
- Seção 3.4 de (Paulino et al., 2003)

6.5.1 Fatores de Bayes

- (Kass and Raftery, 1995)
- Seção 9.5.1 de (Press, 2003)
- Seção 3.4.1 de (Paulino et al., 2003)

6.5.2 FBST - *Full Bayesian Significance Test*

- Proposta de (Pereira and Stern, 1999) para testar hipóteses precisas (*sharp hypotheses*)
- Amplamente revisado em (Pereira et al., 2008) e (Pereira and Stern, 2020)

Chapter 7

Modelos Lineares

A classe de *Modelos Lineares* atende a uma ampla gama de problemas aplicados, apresentada em profundidade por (Neter et al., 2005). Para uma introdução à classe de *Modelos Lineares Generalizados* recomenda-se (McCullagh and Nelder, 1989).

7.1 Correlação

7.2 Regressão Linear Simples

7.2.1 Modelo

O modelo de *Regressão Linear Simples* universal/populacional é construído, pela abordagem clássica, com todos os N pares ordenados do universo, e pode ser descrito pela relação a seguir.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{7.1}$$

onde $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon)$.

Na maioria dos casos práticos trabalha-se com amostras, sendo necessário estimar os valores de β_0 e β_1. O método dos *mínimos quadrados (ordinários)* é utilizado para calcular estas estimativas. O princípio do método é minimizar a soma de quadrado dos erros, i.e.,

$$minimizar \sum_{i=1}^{n} \varepsilon_i^2. \tag{7.2}$$

7.2.2 Estimativas dos parâmetros

Basicamente utiliza-se $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$ da Eq. (7.1) e deriva-se (parcialmente) em relação a β_0 e β_1, fazendo cada uma das derivadas parciais igual a zero. Para maiores detalhes recomenda-se (DeGroot and Schervish, 2012). As estimativas por mínimos quadrados são enfim dadas por

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \tag{7.3}$$

e

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{7.4}$$

7.2.3 Análise de diagnóstico

Teste para β_0

As hipóteses usuais do teste para β_0 são $H_0 : \beta_0 = 0$ *vs* $H_1 : \beta_0 \neq 0$. Sob H_0

$$T_0 = \frac{\hat{\beta}_0}{ep(\hat{\beta}_0)} \sim t_{n-2}, (\#eq : teste_b 0) \tag{7.5}$$

onde

$$ep(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = \sqrt{\frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}{n-2} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=i}^{n}(x_i - \bar{x})^2}\right]}.(\#eq : ep_b 0) \tag{7.6}$$

Teste para β_1

O teste para β_1 é fundamental na análise de diagnóstico. É com ele que decide-se a respeito da presença ou ausência de relação linear entre X e Y. As hipóteses usuais do teste para β_1 são $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$. Sob H_0

$$T_1 = \frac{\hat{\beta}_1}{ep(\hat{\beta}_1)} \sim t_{n-2}, \tag{7.7}$$

onde

$$ep(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=i}^{n}(x_i - \bar{x})^2}}.(\#eq : ep_t 1) \tag{7.8}$$

7.2.4 Modelo RPO

Um modelo na forma da Eq. (7.9) é chamado *regressão pela origem* pelo fato de a reta ajustada passar pelo ponto $(0, 0)$, a *origem* do plano cartesiano.

$$Y_i = \beta_1 X_i + \varepsilon_i, \tag{7.9}$$

onde $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon)$. A estimativa do parâmetro β_1 é dada por

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}.(\#eq : beta1_r po) \tag{7.10}$$

Teste para β_1 do modelo RPO

$$T_1^{RPO} = \frac{\hat{\beta}_1}{ep(\hat{\beta}_1)} \sim t_{n-1}, (\#eq : teste_b eta1_r po) \tag{7.11}$$

onde

$$ep(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2/(n-1)}{\sum_{i=i}^{n} x_i^2}}.(\#eq : ep_b eta1_r po) \tag{7.12}$$

Exemplo 7.1. O dono de um bar decidiu modelar o número de garrafas de bebidas vendidas em seu estabelecimento (Y) em função da temperatura máxima do dia (X).

```
x <- read.table('http://www.filipezabala.com/data/drinks.txt', header = T, sep = '\t')
dim(x)  # dimensão de x
```

```
## [1] 30  2
```

```
head(x) # cabeçalho de x
```

```
##   temp gar
## 1 29.5 145
## 2 31.3 170
## 3 34.7 167
## 4 40.4 244
## 5 28.4 159
## 6 40.3 195
```

```
fit <- lm(gar ~ temp, data = x)       # regressão linear simples
summary(fit)
```

```
##
## Call:
## lm(formula = gar ~ temp, data = x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.16  -8.96   3.58  10.81  33.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -19.334     22.944   -0.84     0.41
## temp           5.920      0.674    8.78  1.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.3 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.734,  Adjusted R-squared:  0.724
## F-statistic: 77.1 on 1 and 28 DF,  p-value: 1.57e-09
```

```
fit0 <- lm(gar ~ temp - 1, data = x)  # regressão pela origem
summary(fit0)
```

```
##
## Call:
## lm(formula = gar ~ temp - 1, data = x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.78 -11.26   3.53  12.01  30.29
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## temp   5.3582     0.0975      55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.2 on 29 degrees of freedom
## Multiple R-squared:  0.99,   Adjusted R-squared:  0.99
## F-statistic: 3.02e+03 on 1 and 29 DF,  p-value: <2e-16
```

Exercício 7.1. Considere o Exemplo 5.1.
(a) Indique os valores de n e p.
(b) No modelo de regressão linear simples atribuído a `fit`, indique os números das equações para o cálculo das medidas `Estimate`, `Std. Error` e `t value`.
(c) Utilizado a linguagem R, aplique as equações indicadas no item (b) de forma a obter os valores indicados em `summary(fit)`.
(d) No modelo de regressão pela origem atribuído a `fit0`, indique os números das equações para o cálculo das medidas `Estimate`, `Std. Error` e `t value`.
(e) Utilizado a linguagem R, aplique as equações indicadas no item (d) de forma a obter os valores indicados em `summary(fit0)`.
(f) Qual o modelo mais indicado segundo a sua análise? Justifique.

7.3 Regressão Linear Múltipla

O modelo de *regressão linear múltipla* universal/populacional é construído, pela abordagem clássica, com todos os N pares ordenados do universo, e pode ser descrito pela relação a seguir.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \tag{7.13}$$

$\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon)$. Por ter dimensionalidade p usualmente utiliza-se notação matricial na forma

$$Y = X\beta + \varepsilon, \tag{7.14}$$

onde $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$ e $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$.

Para a obtenção das estimativas dos parâmetros utiliza-se a Eq. (7.15).

$$\hat{\beta} = (X^{'}X)^{-1}(X^{'}X)Y \tag{7.15}$$

Exemplo 7.2. No site http://archive.ics.uci.edu/ml/datasets/Energy+efficiency está disponível uma análise de energia feita por (Tsanas and Xifara, 2012) usando 12 formas diferentes de construção simuladas no Ecotect. Os edifícios diferem em relação à área envidraçada, à distribuição da área envidraçada e à orientação, entre outros parâmetros. Foram simuladas várias configurações como funções das características acima mencionadas para obter 768 formas de construção. O conjunto de dados detalhado a seguir compreende 768 amostras e 8 características (`X1` a `X8`), com o objetivo de prever duas respostas reais (`Y1` e `Y2`).

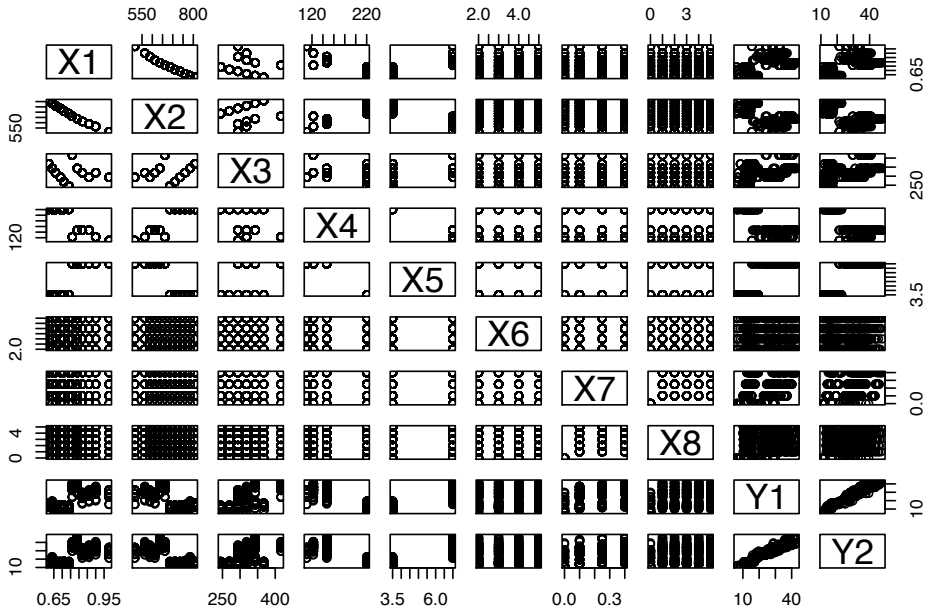
`X1`: Compactação Relativa
`X2`: Superfície
`X3`: Área da parede
`X4`: Área do telhado
`X5`: Altura total
`X6`: Orientação
`X7`: Área de Envidraçamento
`X8`: Distribuição da Área de Envidraçamento
`Y1`: Carga de aquecimento
`Y2`: Carga de resfriamento

```
library(readxl)
url1 <- 'http://archive.ics.uci.edu/ml/machine-learning-databases/00242/ENB2012_data.xlsx'
download.file(url1, 'temp.xlsx', mode = 'wb')
energy <- read_excel('temp.xlsx')
str(energy)   # dando uma olhada nas variáveis
```

```
## tibble [768 x 10] (S3: tbl_df/tbl/data.frame)
##  $ X1: num [1:768] 0.98 0.98 0.98 0.98 0.9 0.9 0.9 0.9 0.86 0.86 ...
##  $ X2: num [1:768] 514 514 514 514 564 ...
##  $ X3: num [1:768] 294 294 294 294 318 ...
##  $ X4: num [1:768] 110 110 110 110 122 ...
##  $ X5: num [1:768] 7 7 7 7 7 7 7 7 7 7 ...
##  $ X6: num [1:768] 2 3 4 5 2 3 4 5 2 3 ...
```

```
##  $ X7: num [1:768] 0 0 0 0 0 0 0 0 0 0 ...
##  $ X8: num [1:768] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Y1: num [1:768] 15.6 15.6 15.6 15.6 20.8 ...
##  $ Y2: num [1:768] 21.3 21.3 21.3 21.3 28.3 ...
```

```
pairs(energy) # matriz de dispersão
```




```
fit0 <- lm(Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = energy) # modelo saturad
summary(fit0)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = energy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.897 -1.320 -0.025  1.353  7.705
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.01342   19.03361    4.41  1.2e-05 ***
## X1          -64.77343   10.28945   -6.30  5.2e-10 ***
## X2           -0.08729    0.01708   -5.11  4.0e-07 ***
## X3            0.06081    0.00665    9.15  < 2e-16 ***
## X4                 NA         NA      NA       NA
## X5            4.16995    0.33799   12.34  < 2e-16 ***
## X6           -0.02333    0.09470   -0.25   0.8055
```

```
## X7           19.93274    0.81399   24.49  < 2e-16 ***
## X8            0.20378    0.06992    2.91   0.0037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.93 on 760 degrees of freedom
## Multiple R-squared:  0.916,  Adjusted R-squared:  0.915
## F-statistic: 1.19e+03 on 7 and 760 DF,  p-value: <2e-16
```

```
fit1 <- step(fit0)  # filtrando as variáveis com stepwise
```

```
## Start:  AIC=1661
## Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##
##
## Step:  AIC=1661
## Y1 ~ X1 + X2 + X3 + X5 + X6 + X7 + X8
##
##        Df Sum of Sq   RSS  AIC
## - X6    1         1  6544 1659
## <none>               6544 1661
## - X8    1        73  6617 1668
## - X2    1       225  6769 1685
## - X1    1       341  6885 1698
## - X3    1       721  7264 1740
## - X5    1      1311  7854 1800
## - X7    1      5163 11707 2106
##
## Step:  AIC=1659
## Y1 ~ X1 + X2 + X3 + X5 + X7 + X8
##
##        Df Sum of Sq   RSS  AIC
## <none>               6544 1659
## - X8    1        73  6617 1666
## - X2    1       225  6769 1683
## - X1    1       341  6886 1697
## - X3    1       721  7265 1738
## - X5    1      1311  7855 1798
## - X7    1      5163 11707 2104
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2 + X3 + X5 + X7 + X8, data = energy)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -9.931 -1.319 -0.026  1.359  7.717
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  83.93176   19.01898    4.41  1.2e-05 ***
## X1          -64.77343   10.28310   -6.30  5.1e-10 ***
## X2           -0.08729    0.01706   -5.12  4.0e-07 ***
## X3            0.06081    0.00664    9.15  < 2e-16 ***
## X5            4.16995    0.33778   12.35  < 2e-16 ***
## X7           19.93274    0.81348   24.50  < 2e-16 ***
## X8            0.20378    0.06987    2.92   0.0036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.93 on 761 degrees of freedom
## Multiple R-squared:  0.916,  Adjusted R-squared:  0.916
## F-statistic: 1.39e+03 on 6 and 761 DF,  p-value: <2e-16
```

Após o primeiro ajuste atribuído a `fit0` é possível notar que o coeficiente da variável `X4` não é possível de ser calculado devido a singularidades, i.e, impossibilidade de inversão das matrizes do modelo. Sendo assim, ao modelo saturado (contendo todas as variáveis candidatas) aplicou-se o método de *stepwise*, proposto por (Efroymson, 1960) e utilizado para selecionar variáveis. Este método busca automaticamente o melhor conjunto de variáveis de maneira a minimizar alguma medida, usualmente o *Critério de Informação de Akaike* (AIC, na sigla em inglês), sugerido por (Akaike, 1974). De acordo com a métrica do stepwise, quanto menor o valor de AIC, melhor a combinação das variáveis.

Pelos resultados obtidos pode-se verificar que o modelo ajustado em `fit1` possui todas as variávies significativas para um α inferior a 0.01, estatística F de 1387 para 6 e 761 graus de liberdade, levando a um p-value geral do modelo menor que 2.2×10^{-16}, o que indica boa aderência aos dados. O valor do `Multiple R-squared` é de 0.9162, indicando que o modelo explica em torno de 92% da variação de `Y1`. Desta forma este é um modelo aceitável, que possui coeficientes de `X1` e `X2` negativos, indicando que um aumento destas variáveis (respectivamente compactação relativa e superfície) deve reduzir a carga de aquecimento. Mais especificamente, um aumento de uma unidade na compactação relativa (`X1`) gera uma redução esperada de 64.77 unidades na carga de aquecimento, mantidas constantes as demais variáveis. As variáveis `X3`, `X5`, `X7` e `X8` possuem coeficientes positivos, levando a um impacto esperado positivo em `Y1`. Como exemplo, para cada aumento de uma unidade na altura total (`X5`) espera-se um aumento aproximando de 4.17 unidades na carga de aquecimento. As outras variáveis possuem interpretação análoga, devendo-se sempre observar o sinal dos coeficientes.

Exercício 7.2. Refaça o Exemplo 5.2 utilizando a variável `Y2` como resposta,

ajustando o modelo saturado, filtrando as variáveis com a função `step` e interpretando os resultados.

7.4 Regressão Logística

7.4.1 Variáveis binárias/dicotômicas

Em problemas aplicados é comum fazer uso de variáveis aleatórias que admitam apenas dois valores, chamadas v.a. *binárias* ou *dicotômicas*. Empresas de serviços financeiros podem estar interessados em clientes adimplentes/inadimpletes, hospitais em pacientes com/sem melhora, cientistas da computação em servidores operantes/inoperantes, etc. Começamos com um exemplo numérico para ilustrar.

Exemplo 7.3. Um banco está interessado na variável aleatória Y: cliente inadimplente. Ela assume valor 1 se o cliente é inadimplente (sucesso) e 0 caso contrário (fracasso). Note que a terminologia sucesso/fracasso indica se a variável Y foi observada (1) ou não (0), ainda que 'sucesso' possa não significar algo agradável.
Se considerarmos 40 clientes, 20 inadimplentes e 20 adimplentes, pode-se calcular a probabilidade (incondicional) de observar um ciente inadimplente (sucesso) por $Pr(Y = 1) = \frac{20}{40} = 0.5$. Da mesma forma, a probabilidade (incondicional) de observar um ciente adimplente (fracasso) é dada por $Pr(Y = 0) = \frac{20}{40} = 0.5$. Utilizando a linguagem R pode-se gerar a sequência de zeros e uns descrita acima, bem como as respectivas probabilidades.

```
n <- 20
(y <- c(rep(0,n), rep(1,n)))
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
(taby <- table(y))
```

```
## y
##  0  1
## 20 20
```

```
prop.table(taby)
```

```
## y
##   0   1
## 0.5 0.5
```

Vendo a questão desta maneira pode-se considerar que a probablidade de um cliente ser inadimplente é de 50%. É possível, porém, considerar outras variáveis para refinar esta probabilidade. Suponha uma variável X_1, que ocorre em aproximadamente 20% dos clientes adimplentes ($Y = 0$) e em aproximadamente

90% dos clientes inadimplentes ($Y = 1$). Assim, visto que presença da variável X_1 é maior entre os inadimplentes, intuitivamente devemos atribuir uma probabilidade maior de inadimplência a clientes que apresentem a característica X_1, e menor para aquele onde a característica X_1 é ausente. Novamente podemos fazer uso da linguagem R para gerar sequências similares àquelas consideradas na teoria.

```
suppressMessages(library(tidyverse))
set.seed(1); (x1 <- c(rbinom(n,1,.2), rbinom(n,1,.9)))  # gerando sequência pseudoalea
```

```
##  [1] 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Note na saída abaixo que o segmento indicado por `$`0`` refere-se ao grupo onde $X_1 = 0$, e `$`1`` onde $X_1 = 1$.

```
by(y,x1,table) %>%   # contando o número de zeros e uns em y separado por x1
  lapply(prop.table) # calculando as proporções
```

```
## $`0`
##
##      0      1
## 0.9412 0.0588
##
## $`1`
##
##     0     1
## 0.174 0.826
```

Note que a probabilidade de inadimplência no grupo onde X_1 não é observada é igual a $Pr(Y = 1|X_1 = 0) = 0.05882353$. No grupo de clientes que apresentam a característica X_1 esta probabilidade sobe para $Pr(Y = 1|X_1 = 1) = 0.826087$.

Finalmente é considerada uma variável X_2, que aparece em aproximadamente metade dos clientes inadimplentes e em aproximadamente metade dos adimplentes. Assim, observar a característica X_2 não deve trazer informação sobre a probabilidade de inadimplência, simbolizada pela probabilidade condicional $Pr(Y = 1|X_2 = x_2)$, $x_2 \in \{0, 1\}$.

```
set.seed(1); (x2 <- rbinom(2*n,1,.5)) # simulando valores de X2
```

```
##  [1] 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 1 1 0 1 0 0 0 0 0 1 0 0 1 0 0 1 1 1 0 1 0
```

```
by(y,x2,table) %>%  # contando os zeros e uns de y, separados pelos valores de X2
  lapply(prop.table) # transformando a contagem em proporção
```

```
## $`0`
##
##     0     1
## 0.429 0.571
##
```

```
## $`1`
##
##     0     1
## 0.579 0.421
```

7.4.2 O modelo de regressão logística

A *regressão logística* pertence à classe dos *modelos lineares generalizados*, descrita em detalhes por (McCullagh and Nelder, 1989), (Agresti, 2007) e (Paula, 2013). Seja Y uma variável aleatória binária com distribuição binimial de probablidade de sucesso $\pi(x)$. A notação $\pi(x)$ sugere que a probabilidade de sucesso está condicionada a um valor/categoria x. Desta forma, $\pi(x) = Pr(Y = 1|X = x)$. Define-se a função *logito* conforme a Eq. @ref(eq:logit_uni), onde log indica o logaritmo na base $e \approx 2.718281828459$.

$$\text{logito}\left[\pi(x)\right] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x (\#eq:logit_uni) \tag{7.16}$$

Isolando $\pi(x)$ na Eq. @ref(eq:logit_uni) obtém-se

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} (\#eq:pix_uni) \tag{7.17}$$

Exemplo 7.4. Considerando novamente as informações do Exemplo 8.1, vamos agora utilizar a estrutura das Equações @ref(eq:logit_uni) e @ref(eq:pix_uni) para abordar o problema.

```
# modelo 1
x1 <- as.factor(x1)  # convertendo em fator para usar a função glm
fit1 <- glm(y ~ x1, family = 'binomial') # ajustando modelo logístico
summary(fit1) # detalhamento do modelo
```

```
##
## Call:
## glm(formula = y ~ x1, family = "binomial")
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.870  -0.348   0.135   0.618   2.380
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.77       1.03   -2.69  0.00715 **
## x11             4.33       1.17    3.71  0.00021 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55.452  on 39  degrees of freedom
## Residual deviance: 28.860  on 38  degrees of freedom
## AIC: 32.86
##
## Number of Fisher Scoring iterations: 5
```

Note que o intercepto $\widehat{\beta}_0 = -2.773$ é significativo ($p - value = 0.00715$), bem como o coeficiente que indica a presença do atributo X_1, $\widehat{\beta}_1 = 4.331$ ($p-value = 0.00021$). Assim, pode-se considerar o modelo conforme Eq. @ref(eq:logit_uni), na forma

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = -2.773 + 4.331 I_{x_1}.$$

A simbologia I_{x_1} indica uma variável *indicadora* da presença do atributo X_1. Assim, se a pessoa possuir o atributo X_1 considera-se $I_{x_1} = 1$, e $I_{x_1} = 0$ caso contrário. Este é um modelo conhecido como *casela de referência*, visto que a variável X_1 é categórica. Desta forma, uma das categorias/níveis da variável é escolhida como a (casela de) referência. Por padrão, o R utiliza a primeira da ordem numérica/alfabética, no caso $X_1 = 0$. Desta forma, se a pessoa não possui a característica X_1, pode-se calcular sua probabilidade de inadimplência pela Eq. @ref(eq:pix_uni), dada por

$$Pr(Y = 1|X_1 = 0) = \pi(0) = \frac{e^{-2.773+4.331\times 0}}{1 + e^{-2.773+4.331\times 0}} \approx 0.05882353.$$

No caso de alguém que possui a característica X_1,

$$Pr(Y = 1|X_1 = 1) = \pi(1) = \frac{e^{-2.773+4.331\times 1}}{1 + e^{-2.773+4.331\times 1}} \approx 0.826087.$$

```
# modelo 1
exp(coef(fit1)[1])/(1+exp(coef(fit1)[1])) # Pr(Y = 1 | X1 = 0)
```

```
## (Intercept)
##      0.0588
```

```
exp(sum(coef(fit1)))/(1+exp(sum(coef(fit1)))) # Pr(Y = 1 | X1 = 1)
```

```
## [1] 0.826
```

O mesmo procedimento considerado para X_1 pode ser realizado para X_2. Note que o intercepto $\widehat{\beta}_0 = 0.2877$ é **não** significativo ($p - value = 0.514$), bem como o coeficiente que indica a presença do atributo X_2, $\widehat{\beta}_1 = -0.6061$ ($p - value = 0.344$).

```r
# modelo 2
x2 <- as.factor(x2)  # convertendo em fator para usar a função glm
fit2 <- glm(y ~ x2, family = 'binomial') # ajustando modelo logístico
summary(fit2) # detalhamento do modelo
```

```
##
## Call:
## glm(formula = y ~ x2, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3018  -1.0455   0.0062   1.0579   1.3153
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.288      0.441    0.65     0.51
## x21           -0.606      0.641   -0.95     0.34
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55.452  on 39  degrees of freedom
## Residual deviance: 54.546  on 38  degrees of freedom
## AIC: 58.55
##
## Number of Fisher Scoring iterations: 4
```

Desta maneira o modelo não deve ser utilizado, mas para efeito de comparação com os resultados do Exemplo 8.1 são calculadas as probabilidades de sucesso condicionadas a $X_2 = 0$ e $X_2 = 1$.

```r
# modelo 2
exp(coef(fit2)[1])/(1+exp(coef(fit2)[1])) # Pr(Y = 1 | X2 = 0)
```

```
## (Intercept)
##       0.571
```

```r
exp(sum(coef(fit2)))/(1+exp(sum(coef(fit2)))) # Pr(Y = 1 | X2 = 1)
```

```
## [1] 0.421
```

Exercício 7.3. Considere o conjunto de dados apresentado por Ronny Kohavi e Barry Becker, disponível em https://archive.ics.uci.edu/ml/datasets/adult. Considere um modelo de regressão logística para avaliar as características que mais impactam no salario das pessoas (acima ou abaixo de 50 mil dólares).

```r
# lendo e arrumando os dados
dat <- read.table('https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data',
```

```
                    sep = ',')
dat <- dat[,-c(3,9)]
colnames(dat) <- c('idade','tipoTrabalho','educacao','anosEstudo','estadoCivil','ocupa
                   'relacao','genero','ganhoCapital','perdaCapital','horasPorSemana',
                   'paisOrigem','salario')
```

Chapter 8

Aprendizado de Máquina

A aplicação de *aprendizado de máquina* ou *modelagem algorítmica* está em crescente expansão. A aplicação deste tipo de metodologia usualmente utiliza *modelagem preditiva*, e não inferencial. Para maiores detalhes veja a discussão nas Seções 1.2 e 1.3 de (Izbicki and dos Santos, 2020), bem como nos slides disponibilizados pelo professor neste link.

8.1 Análise de Componentes Principais (*PCA*)

A *Análise de Componentes Principais* (*PCA*, na sigla em inglês) é uma técnica de redução de dimensionalidade usualmente aplicada a um grande número de variáveis relacionadas, de forma a capturar o máximo possível da variabilidade do conjunto de dados. Foi introduzida por (Pearson, 1901) e estudada independentemente por (Hotelling, 1933) e outros pesquisadores que abordaram o problema de formas variadas. Considerando a definição de (Bishop, 1999), seja um conjunto de dados X de dimensão $n \times p$ composto por n vetores p-dimensionais conforme indicado a seguir.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Deste conjunto de dados calcula-se a matriz de covariâncias amostrais S dada por

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T, \tag{8.1}$$

onde $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ é o vetor de médias amostrais. São obtidos os autovetores

u_i e os autovalores λ_i de S pela equação

$$Su_i = \lambda_i u_i, \tag{8.2}$$

$i = 1, \ldots, p$. Os autovetores correspondentes aos q maiores autovetores ($q < p$) são mantidos, e uma representação de dimensão reduzida é definida por uma combinação linear dos autovetores e dos dados deslocados pela média. Matematicamente, $d_n = U^T(x_n - \bar{x})^T$ onde $U_q = (u_1, \ldots, u_q)$.

```
df <- iris[-5]      # retirando a quinta coluna, 'Species'
(m <- colMeans(df)) # vetor de médias
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##         5.84         3.06         3.76         1.20
```

```
(S <- cov(df))      # matriz de covariâncias
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length       0.6857     -0.0424         1.27       0.516
## Sepal.Width       -0.0424      0.1900        -0.33      -0.122
## Petal.Length       1.2743     -0.3297         3.12       1.296
## Petal.Width        0.5163     -0.1216         1.30       0.581
```

```
eigen(S)            # autovalores (variâncias) e autovetores de S
```

```
## eigen() decomposition
## $values
## [1] 4.2282 0.2427 0.0782 0.0238
##
## $vectors
##         [,1]    [,2]    [,3]   [,4]
## [1,]  0.3614 -0.6566 -0.5820  0.315
## [2,] -0.0845 -0.7302  0.5979 -0.320
## [3,]  0.8567  0.1734  0.0762 -0.480
## [4,]  0.3583  0.0755  0.5458  0.754
```

```
(av <- prcomp(df))  # via função
```

```
## Standard deviations (1, .., p=4):
## [1] 2.056 0.493 0.280 0.154
##
## Rotation (n x k) = (4 x 4):
##                    PC1     PC2     PC3    PC4
## Sepal.Length  0.3614 -0.6566  0.5820  0.315
## Sepal.Width  -0.0845 -0.7302 -0.5979 -0.320
## Petal.Length  0.8567  0.1734 -0.0762 -0.480
## Petal.Width   0.3583  0.0755 -0.5458  0.754
```

É possível realizar o mesmo procedimento na matriz de correlação R. Esta abordagem é recomendada para evitar que os resultados sejam afetados pela

escala dos valores observados.

```
(R <- cor(df))  # matriz de correlação
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length        1.000      -0.118        0.872       0.818
## Sepal.Width        -0.118       1.000       -0.428      -0.366
## Petal.Length        0.872      -0.428        1.000       0.963
## Petal.Width         0.818      -0.366        0.963       1.000
```

```
eigen(R)        # autovalores e autovetores de R
```

```
## eigen() decomposition
## $values
## [1] 2.9185 0.9140 0.1468 0.0207
##
## $vectors
##        [,1]    [,2]   [,3]   [,4]
## [1,]  0.521 -0.3774  0.720  0.261
## [2,] -0.269 -0.9233 -0.244 -0.124
## [3,]  0.580 -0.0245 -0.142 -0.801
## [4,]  0.565 -0.0669 -0.634  0.524
```

```
prcomp(df, scale = T)  # via função, scale = TRUE
```

```
## Standard deviations (1, .., p=4):
## [1] 1.708 0.956 0.383 0.144
##
## Rotation (n x k) = (4 x 4):
##                    PC1     PC2    PC3    PC4
## Sepal.Length  0.521 -0.3774  0.720  0.261
## Sepal.Width  -0.269 -0.9233 -0.244 -0.124
## Petal.Length  0.580 -0.0245 -0.142 -0.801
## Petal.Width   0.565 -0.0669 -0.634  0.524
```

A proporção da variância explicada pelo i-ésimo componente principal é dada pela Eq. (8.3), e pode ser visualizada em um gráfico ordenado, usualmente chamado *screeplot*.

$$PVE_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{8.3}$$

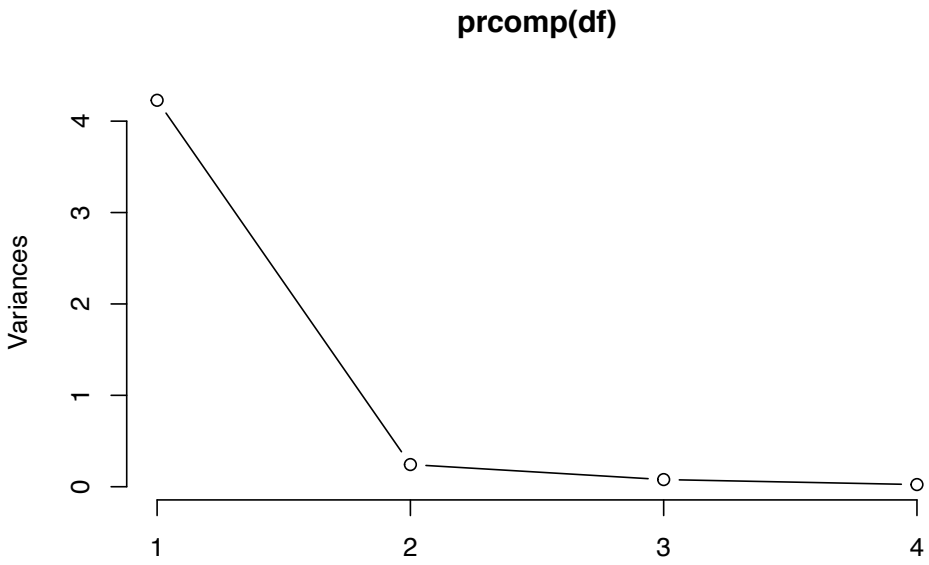
```
(vS <- eigen(S)$values) # autovalores (variâncias) a partir de S
```

```
## [1] 4.2282 0.2427 0.0782 0.0238
```

```
vS/sum(vS)  # Equação (17)
```

```
## [1] 0.92462 0.05307 0.01710 0.00521
```

```
screeplot(prcomp(df), type = 'lines')
```



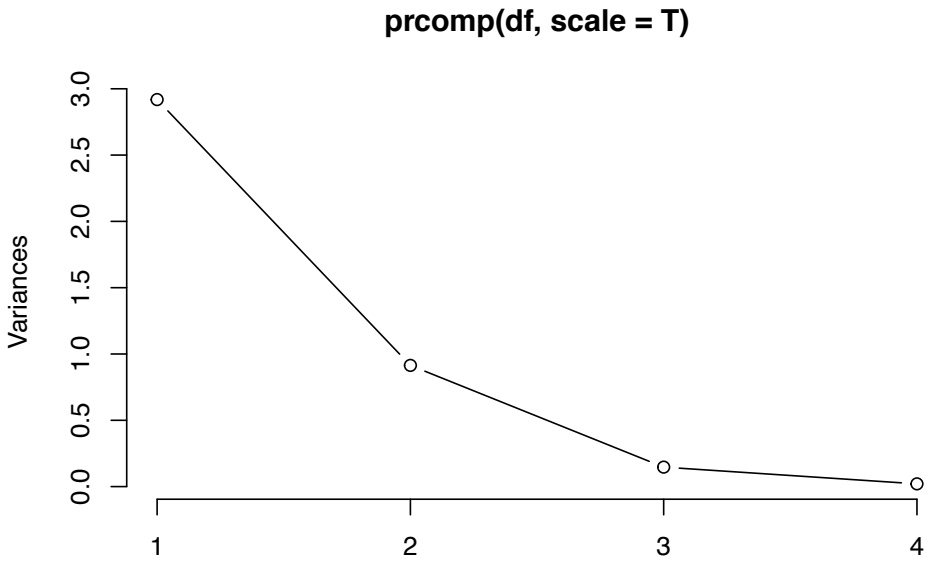

```
(vR <- eigen(R)$values) # autovalores (variâncias) a partir de R
```

```
## [1] 2.9185 0.9140 0.1468 0.0207
```

```
vR/sum(vR)  # Equação (17)
```

```
## [1] 0.72962 0.22851 0.03669 0.00518
```

```
screeplot(prcomp(df, scale = T), type = 'lines')
```



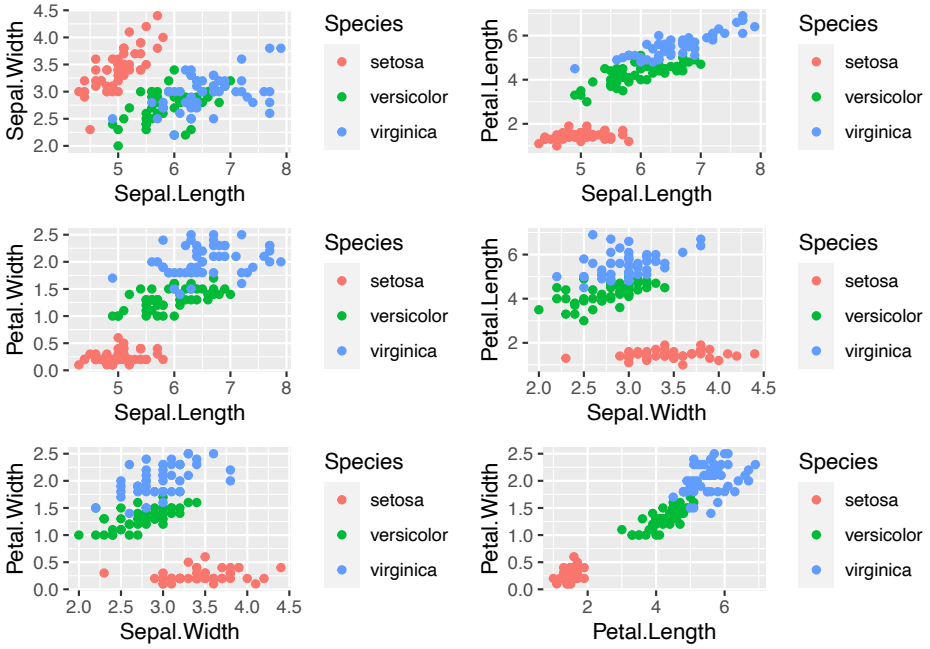
Considere o banco de dados `iris`, que contém 4 colunas numéricas com as larguras e comprimentos das pétalas e sépalas de três espécies de flores do gênero íris.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 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
```

Existem $\binom{4}{2} = 6$ combinações possíveis de gráficos bidimensionais, apresentados a seguir.

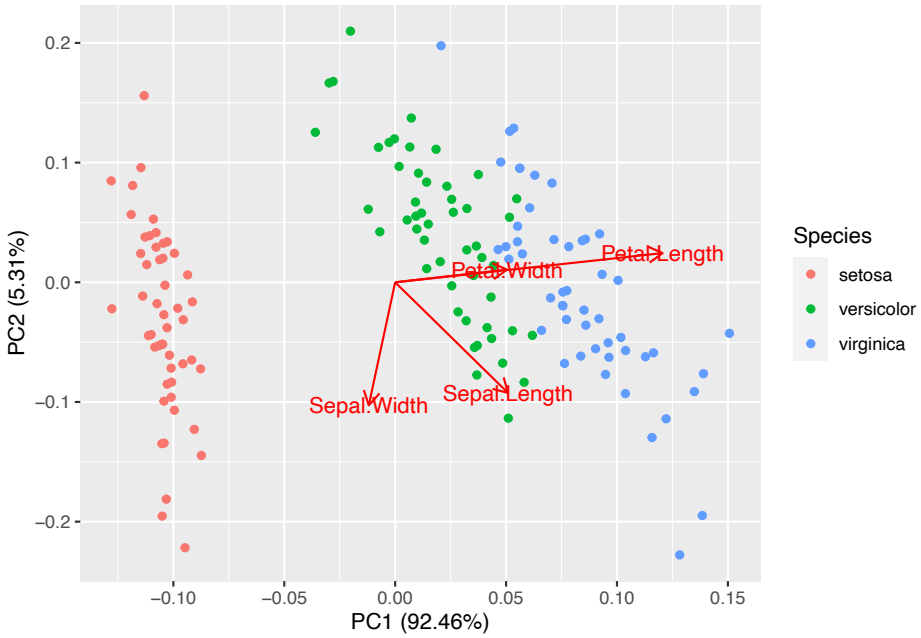
```
require(gridExtra)
p1 <- ggplot(iris, aes(Sepal.Length, Sepal.Width, colour = Species)) + geom_point()
p2 <- ggplot(iris, aes(Sepal.Length, Petal.Length, colour = Species)) + geom_point()
p3 <- ggplot(iris, aes(Sepal.Length, Petal.Width, colour = Species)) + geom_point()
p4 <- ggplot(iris, aes(Sepal.Width, Petal.Length, colour = Species)) + geom_point()
p5 <- ggplot(iris, aes(Sepal.Width, Petal.Width, colour = Species)) + geom_point()
p6 <- ggplot(iris, aes(Petal.Length, Petal.Width, colour = Species)) + geom_point()
grid.arrange(p1, p2, p3, p4, p5, p6, ncol = 2)
```



É posível utilizar o método de componentes principais para aprimorar a visual-

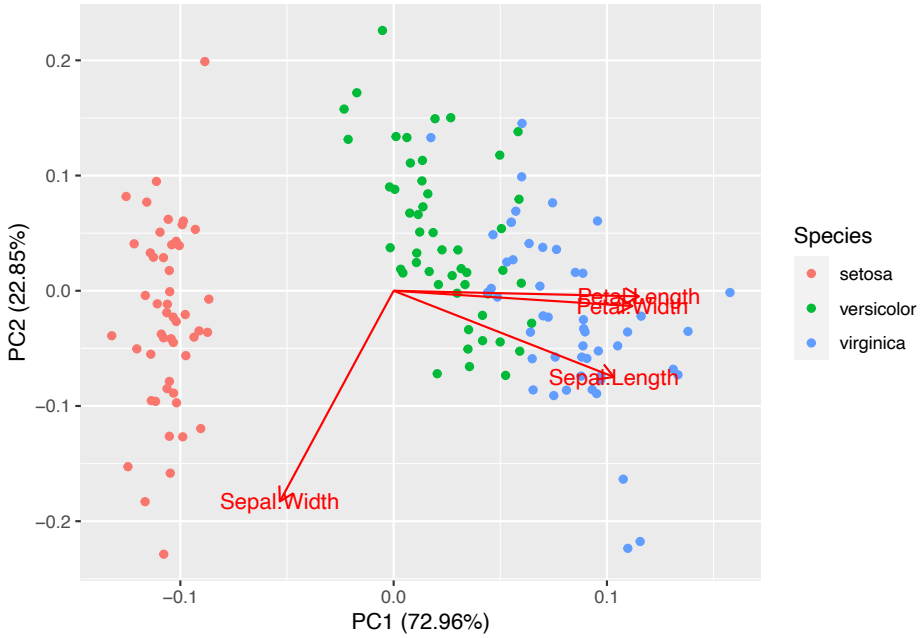
ização[1] da estrutura de associação entre as diferentes espécies de plantas.

```
library(ggfortify)
autoplot(prcomp(df), data = iris, colour = 'Species', loadings = T, loadings.label = T
```




```
autoplot(prcomp(df, scale = T), data = iris, colour = 'Species', loadings = T, loadings
```

[1] Exemplos baseados em https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html.



Exercício 8.1. Considere o banco de dados sobre câncer de mama apresentado por (Dua and Graff, 2019). A partir do código abaixo, faça uma anáise de componentes principais desconsiderando as duas primeiras colunas, que indicam respectivamente o código de identificação da paciente (`V1`) e o diagnóstico (`V2`, Benigno/Maligno).
(a) Quais os valores de n e p?
(b) O que ocorre no comando `x2 <- x[,-c(1,2)]`?
(c) Obtenha os autovalores e autovetores.
(d) Apresente o *screeplot*.
(e) Apresente o gráfico dos dois primeiros componentes principais colorido por `V2`.
(f) Você considera que é possível associar os diagnósticos às variáveis `V3` a `V32`? Dica: observe se há algum tipo de agrupamento no gráfico do item (e).
(g) Quais variáveis mais influenciam nos compomentes principais 1 e 2? Dica: use `loadings.label = T` na função `autoplot` do item (e) e observe o gráfico.

```
library(ggfortify)
x <- read.table('https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsi
x2 <- x[,-c(1,2)]
```

8.2 Técnicas de Agrupamento

Seguindo a definição de (Hartigan, 1975), *agrupamento* - *clustering* ou ainda *segmentação de dados* - é o *agrupamento de objetos similares*. Objetiva agregar

observações que sejam similares em relação a características admitidas nos modelos considerados. Podem ser *aglomerativas*, quando definem uma delimitação ascendente - onde cada observação inicia como um grupo e se agrega com outras ao longo das iterações - ou *divisivas* se o cercamento é descendente - quando todas as observações começam em um grupo que vai sendo dividido a cada etapa. A linguagem R possui diversas funções para análise de agrupamento, sendo as principais discutidas neste capítulo. Para informações atualizadas, veja https://cran.r-project.org/web/views/Cluster.html.

8.2.1 Medidas de similaridade e dissimilaridade

Distâncias e *divergências* são métricas utilizadas em problemas de classificação, agrupamento e reconhecimento de padrões. São utilizadas para medir a *similaridade* ou *dissimilaridade* entre pontos, vetores e distribuições. É comum realizar a *padronização*, i.e., subtrair cada valor da média e dividir pelo desvio padrão da coluna à qual o valor pertence. Este procedimento pode ser realizado através da função `base::scale`.

> Medida de similaridade avalia o quão similares são dois entes, ficando entre 0 (sem similariadade) e 1 (completamente similares).
>
> Medida de dissimilaridade indica o quão distintos são dois entes, ficando entre 0 (iguais) e infinito (diferentes).

As seguir são apresentadas algumas das principais distâncias da literatura – enquadradas na definição de medidas de dissimilaridade – e calculadas entre dois vetores x e y, usualmente linhas de uma matriz numérica. Apresenta-se ainda um pequeno banco de dados para a aplicação dos exemplos.

```
# criando data frame 'df' para os exemplos a seguir
df <- data.frame(V1=c(3,1), V2=c(2,4))  # vetores V1 e V2
rownames(df) <- c('x','y')  # rótulo das linhas
df
```

```
##   V1 V2
## x  3  2
## y  1  4
```

```
# padronizando os dados
df.s <- scale(df)
df.s
```

```
##       V1     V2
## x  0.707 -0.707
## y -0.707  0.707
## attr(,"scaled:center")
## V1 V2
##  2  3
## attr(,"scaled:scale")
```

```
##   V1   V2
## 1.41 1.41
```

Distância de Manhattan

A *distância de Manhattan*, *norma 1* ou L_1 é uma medida de dissimilaridade que avalia a distância absoluta entre dois vetores, dada pela Equação (8.4).

$$L_1 = \sum_{i=1}^{n} |x_i - y_i| \tag{8.4}$$

```
sum(abs(df[1,]-df[2,]))  # distância manhattan aplicando Eq. (18)
```

```
## [1] 4
```

```
dist(df, method = 'manhattan')  # distância manhattan via 'dist'
```

```
##   x
## y 4
```

```
dist(df.s, method = 'manhattan')  # distância manhattan via 'dist' dos valores padronizados
```

```
##      x
## y 2.83
```

Distância euclidiana

Uma das mais utilizadas medidas de dissimilaridade da literatura, a *distância euclidiana*, *norma 2* ou L_2 é dada pela Equação (8.5).

$$L_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{8.5}$$

```
sqrt(sum((df[1,]-df[2,])^2))  # distância euclidiana aplicando Eq. (19)
```

```
## [1] 2.83
```

```
dist(df, method = 'euclidean')  # distância euclidiana via 'dist'
```

```
##      x
## y 2.83
```

```
dist(df.s, method = 'euclidean')  # distância euclidiana via 'dist' dos valores padronizados
```

```
##   x
## y 2
```

Distância de Minkowski

A *distância de Minkowski*, *norma p* ou L_p é uma medida de dissimilaridade que generaliza as distâncias de Manhattan e euclidiana. É dada pela Equação (8.6).

$$L_p = \sqrt[p]{\sum_{i=1}^{n}(|x_i - y_i|)^p} \tag{8.6}$$

```
sum((abs(df[1,]-df[2,]))^5)^(1/5)     # distância de Minkowski com p=5 aplicando Eq. (
```

```
## [1] 2.3
```

```
dist(df, method = 'minkowski', p = 5) # distância de Minkowski com p=5 via 'dist'
```

```
##     x
## y 2.3
```

```
dist(df.s, method = 'minkowski', p = 5) # dist. de Minkowski com p=5 via 'dist' dos va
```

```
##      x
## y 1.62
```

Exercício 8.2. Considere a função `stats::dist`.
(a) Verifique sua documentação, fazendo `?dist`.
(b) Compare as distâncias euclidiana e de Minkowski com $p = 2$. O que você observa?
(c) Compare as distâncias de Manhattan e de Minkowski com $p = 1$. O que você observa?

Exercício 8.3. Considere as distâncias da Seção 7.1 aplicada às colunas numéricas do banco de dados `pib`, obtido pelo código abaixo.
(a) Padronize os dados e atribua a uma variável chamada `pib.s`.
(b) Realize os cálculos 'a mão' como nos exemplos, tanto para `pib` quanto para `pib.s`.
(c) Realize novamente os cálculos do item (b) utilizando a função `dist`.

```
pib <- read.table('http://www.filipezabala.com/data/pib.txt', head = T, sep = '\t')
```

8.3 Métodos hierárquicos

Como o nome sugere, os *métodos hierárquicos* buscam uma estrutura hierárquica dos grupos. Esta estrutura geralmente se dá em forma de árvore, onde os objetos são apresentados individualmente como um grupo unitário (folha/*leaf*) que se aglomeram por similaridade em grupos maiores (nós/*nodes*) ligados por um grande grupo que une todos elementos (raiz/*root*). Os passos para realizar um agrupamento hierárquico aglomerativo estão descritos a seguir.

ALGORITMO DE AGRUPAMENTO HIERÁRQUICO AGLOMERATIVO

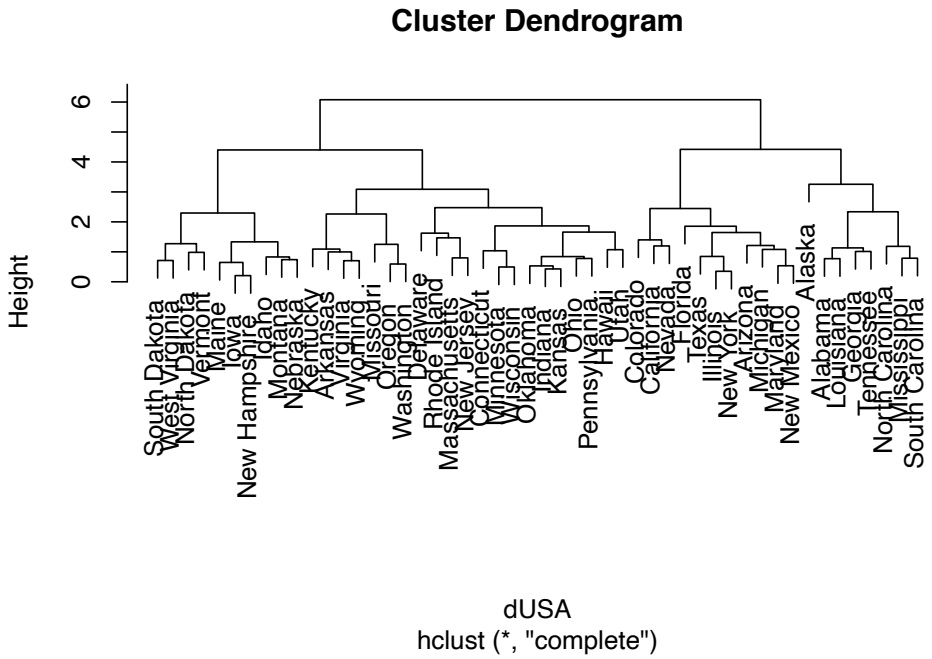
PASSO 1 Padronizar os dados, geralmente com o uso da função `scale`.

PASSO 2 Calcular a (dis)similaridade entre cada par de objetos no conjunto de dados.

PASSO 3 Usar a função de ligação para agrupar os objetos na árvore a partir das informações de distância obtidas na passo 1.

PASSO 4 Apresentar o gráfico da árvore hierárquica em grupos (dendograma), criando uma partição dos dados.

```
# 1. padronizando os dados
USArrest.scale <- scale(USArrests)
# 2. calculando distâncias (utilizando o padrão: euclidean)
dUSA <- dist(USArrest.scale)
# 3. aplicando a função de ligação (utilizando o padrão: complete)
hc <- hclust(dUSA)
# 4. apresentando o gráfico
plot(hc)
```



Os valores no eixo y, intitulados *height*, são as chamadas *distâncias cofenéticas* propostas por (Sokal and Rohlf, 1962). O nome vem da área da Biologia chamada fenética, que estuda métodos de classificação por similaridade

fenotípica. Seu cálculo não é complexo, mas pode ser trabalhoso; assim, será considerada a função `cophenetic` para a obtenção de tais distâncias. Quanto maior for seu valor, mais dissimilar são os elementos comparados. Correlação elevada entre as distâncias calculadas e as distâncias cofenéticas sugere um bom agrupamento.

```
# Calcula a matriz de distâncias cofenéticas
coph <- cophenetic(hc)
sort(unique(coph))
```

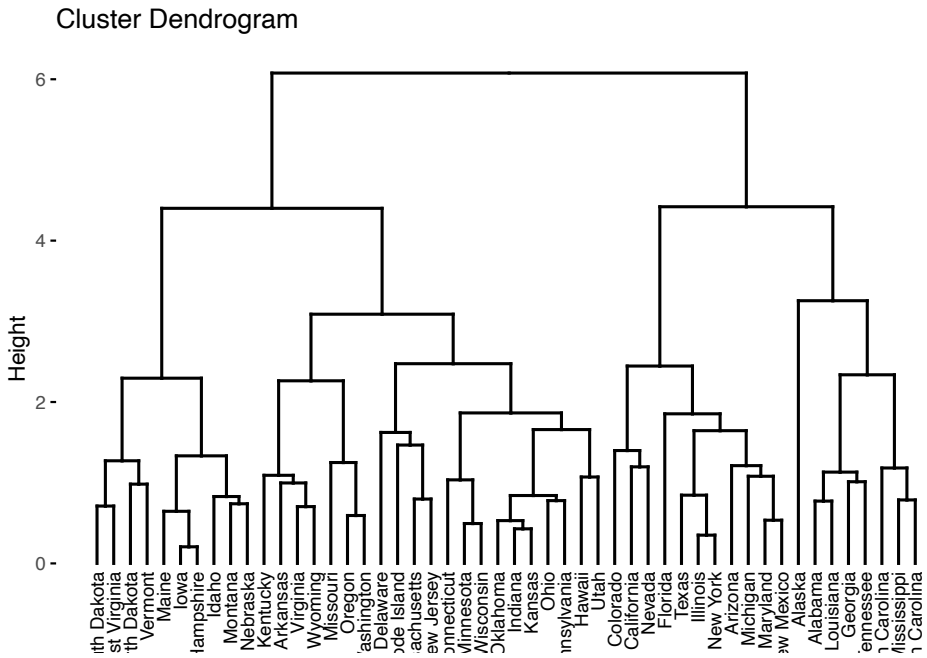
```
##  [1] 0.206 0.350 0.429 0.494 0.530 0.535 0.594 0.646 0.704 0.711 0.739 0.772 0.778 0
## [21] 1.012 1.035 1.071 1.080 1.092 1.131 1.183 1.197 1.212 1.250 1.272 1.333 1.399 1
## [41] 2.295 2.337 2.446 2.475 3.088 3.255 4.401 4.420 6.077
```

```
# Correlação entre as distâncias cofenéticas e as distâncias originais (maior, melhor)
cor(coph,dUSA)
```

```
## [1] 0.698
```

É possível melhorar a visualização do dendograma utilizando a função `factoextra::fviz_dend`.

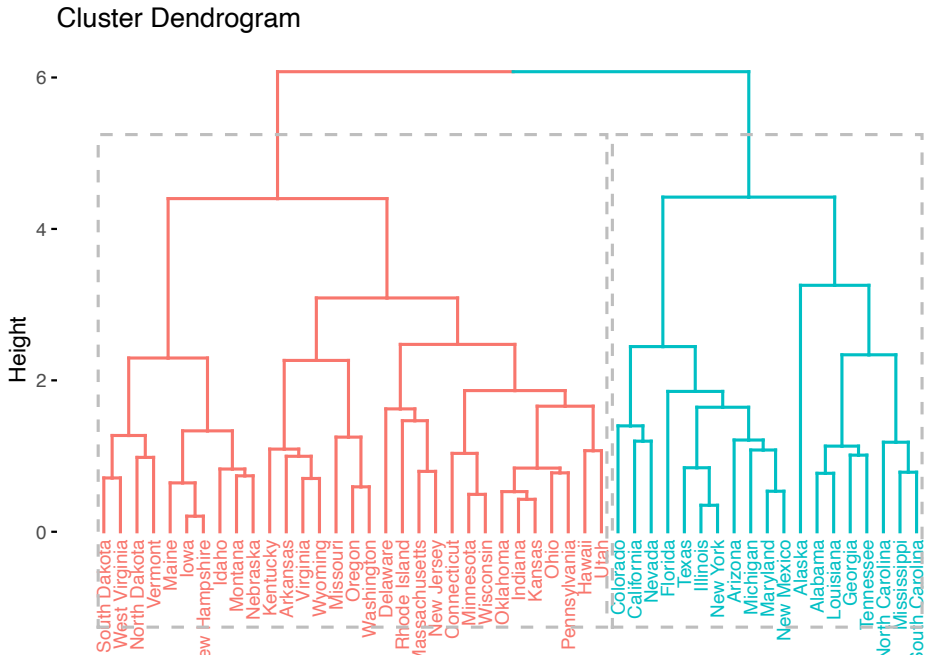
```
library(factoextra)
fviz_dend(hc, cex = 0.6)  # fonte com 60% do tamanho
```



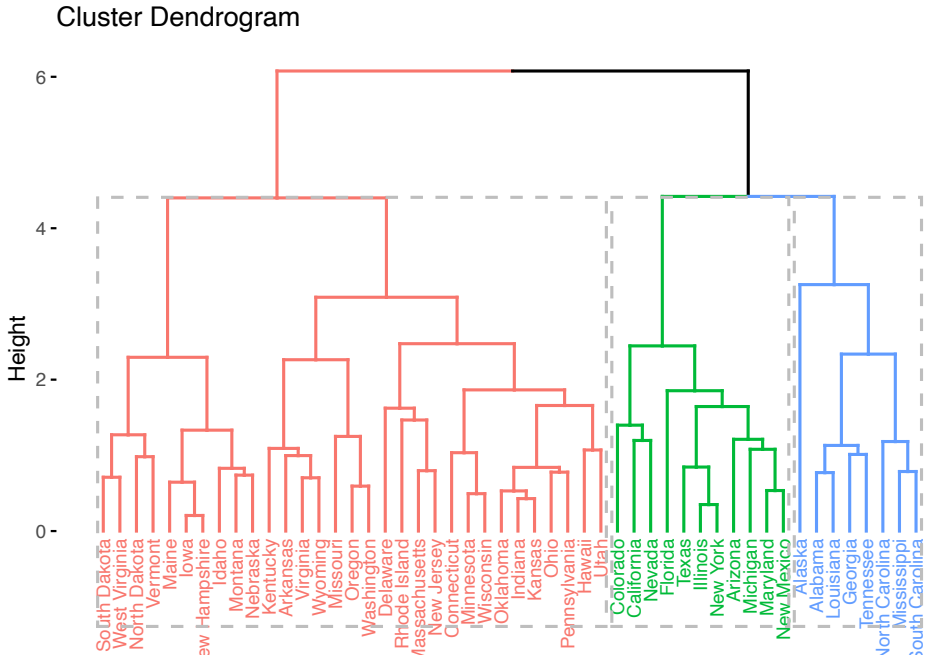
Pode-se utilizar a função `fviz_dend` para colorir um número arbitrário de grupos. Note que os grupamentos são obtidos de cima pra baixo, dependente dos

valores de *height* (distâncias cofenéticas).

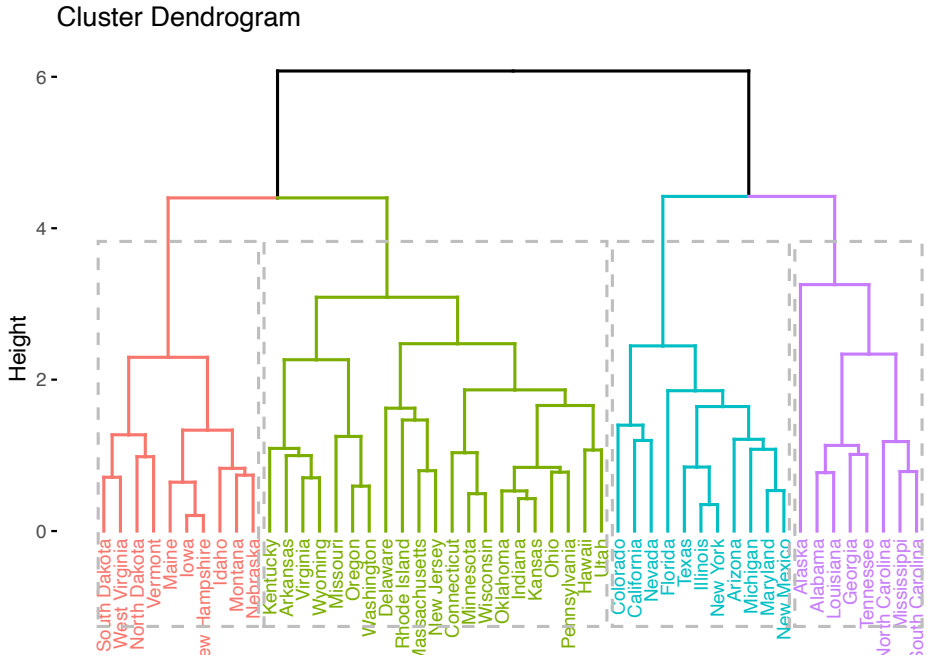
```
fviz_dend(hc, k = 2,  # 2 grupos
          cex = 0.6,  # tamanho do texto/rótulo (label)
          rect = TRUE # adiciona retângulos ao redor dos grupos
)
```




```
fviz_dend(hc, k = 3,  # 3 grupos
          cex = 0.6,  # tamanho do texto/rótulo (label)
          rect = TRUE # adiciona retângulos ao redor dos grupos
)
```




```r
fviz_dend(hc, k = 4,  # 4 grupos
          cex = 0.6,  # tamanho do texto/rótulo (label)
          rect = TRUE # adiciona retângulos ao redor dos grupos
)
```



Exercício 8.4. Considere novamente o conjunto de dados `pib`. Crie dendogramas com a função `fviz_dend` utilizando:
(a) dados originais e padronizados;
(b) diferentes distâncias;
(c) diferentes números de grupos.

8.4 Métodos não hierárquicos (de particionamento)

8.4.1 K-médias

K-médias (*k-means*) é um nome genérico para métodos derivados dos algoritmos de (Lloyd, 1957), (Forgy, 1965), (MacQueen et al., 1967), (Hartigan, 1975) e (Hartigan and Wong, 1979). A ideia básica é encontrar grupos similares, de maneira a minimizar a soma de distâncias euclidianas ao quadrado. As distâncias são calculadas entre os pontos e as médias de cada um dos k grupos, chamadas *centróides*.

Em relação ao modo de busca podem ser classificados como algoritmos de *comutação*, em que objetos devem ser particionados em k de grupos. Uma partição inicial é dada de forma arbitrária, onde se definem k centróides. Calcula-se a distância euclidiana ao quadrado entre as observações e os k centróides. O centróide mais próximo define o grupo ao qual uma observação pertence. Recalculam-se os novos centróides, e novas partições são obtidas com a alternância dos objetos

entre os grupos. O algoritmo encerra quando nenhuma comutação adicional reduz a soma de quadrados intra-grupo, ou quando outro critério de parada é atingido.

São algoritmos relativamente rápidos na execução, mas são afetados pela incerteza da partição inicial. Há sempre a possibilidade de que partições iniciais distintas possam levar a partições finais superiores a outras.

A variação quadrática intra-grupo (VQI_j) do j-ésimo grupo é dada pela Equação (8.7).

$$VQI_j = \sum_{x_i \in G_j} (x_i - \mu_j)^2 \tag{8.7}$$

- x_i é o i-ésimo elemento pertencente ao grupo G_j
- μ_j é o ponto médio do grupo G_j
- $j \in \{2, \ldots, k\}$

A soma de quadrados total (SQT) é dada pela Equação (8.8).

$$SQT = \sum_{i=1}^{k} VQI_i \tag{8.8}$$

Cada observação x_i é atribuída a um grupo de forma que a SQT seja mínima a cada iteração. É recomendado que seja feita a padronização dos dados, de maneira a controlar o impacto da escala na definição dos grupos.

ALGORITMO DAS K-MÉDIAS

PASSO 1 Especifique o número k de grupos a serem criados.

PASSO 2 Selecione arbitrariamente k pontos como centros dos grupos (centróides).

PASSO 3 Atribua cada observação ao grupo de centróide mais próximo, baseado na distância euclidiana entre a observação e os centróides.

PASSO 4 Recalcule os centróides com os pontos atribuídos a cada grupo. O centróide do j-ésimo grupo é um vetor de comprimento p contendo as médias das p variáveis, calculadas com todos os pontos atribuídos ao j-ésimo grupo.

PASSO 5 Repita os passos 3 e 4 até que as atribuições não mais reduzam a soma de quadrados intra-grupo, ou que o número máximo de iterações (ou qualquer outro critério de parada) seja atingido.

Seleção inicial dos centróides

(Hartigan, 1975) sugere que a seleção inicial dos centróides seja baseada na soma dos casos S, que tem um valor mínimo $minS$ e um máximo $maxS$. Para obter k grupos iniciais, propõe atribuir o i-ésimo caso ao j-ésimo grupo, onde j é a parte inteira de

$$k\left(\frac{S - minS}{maxS - minS}\right) + 1 \tag{8.9}$$

Uma adaptação será feita, multiplicando 1.01 a $maxS$ para evitar encontrar $j > k$.

```
iris2 <- scale(iris[-5])
S <- rowSums(iris2)
k <- 2
zab <- k*(S - min(S))/(1.01*max(S)-min(S)) + 1 # atribuição de Zabala (2019) baseada em Hartigan
(g <- floor(zab)) # grupos
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [60] 1 1 2 1 2 1 2 1 1 1 1 2 1 1 1 2 2 2 2 2 1 1 1 1 2 1 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 2
## [119] 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

```
table(g)
```

```
## g
##  1  2
## 83 67
```

```
(centroide <- by(iris2, g, colMeans))
```

```
## INDICES: 1
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##      -0.7071       0.0398      -0.6881      -0.7025
## ------------------------------------------------------------------------------------------
## INDICES: 2
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##       0.8759      -0.0494       0.8524       0.8703
```

```
summary(iris2)
```

```
##   Sepal.Length     Sepal.Width      Petal.Length     Petal.Width
##  Min.   :-1.864   Min.   :-2.426   Min.   :-1.562   Min.   :-1.442
##  1st Qu.:-0.898   1st Qu.:-0.590   1st Qu.:-1.222   1st Qu.:-1.180
##  Median :-0.052   Median :-0.132   Median : 0.335   Median : 0.132
##  Mean   : 0.000   Mean   : 0.000   Mean   : 0.000   Mean   : 0.000
##  3rd Qu.: 0.672   3rd Qu.: 0.557   3rd Qu.: 0.760   3rd Qu.: 0.788
##  Max.   : 2.484   Max.   : 3.080   Max.   : 1.780   Max.   : 1.706
```

Exercício 8.5. Utilize a Equação (23) sem a correção de Zabala e observe o resultado em `iris2`.

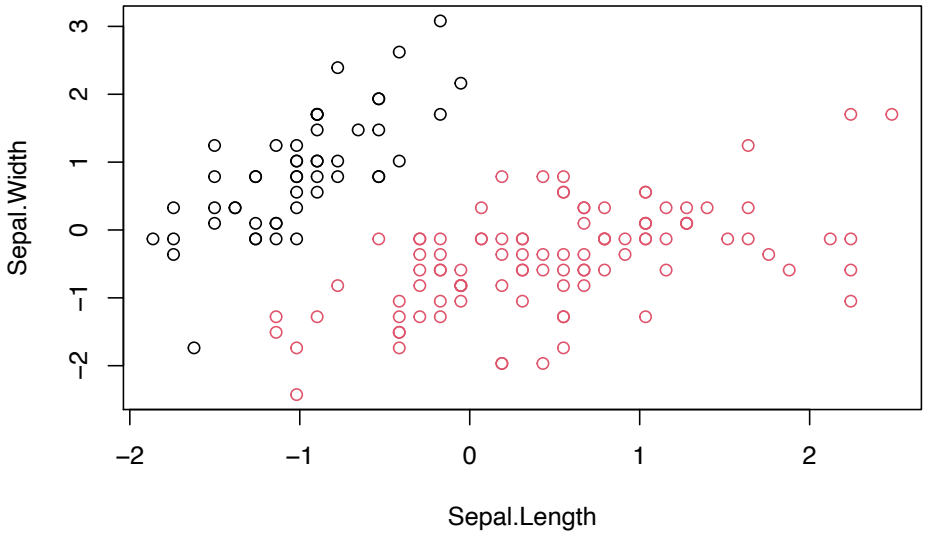
Exercício 8.6. Utilize a Equação (23) com a correção de Zabala para criar uma função que defina os centróides iniciais para um valor genérico k de grupos em uma base de dados numérica. Teste em `iris2` e outros bancos de dados já trabalhados.

Implementando no R

No R pode-se utilizar a função `stats::kmeans` para definir os grupamentos através das k-médias. Por padrão, esta função utiliza 10 como valor padrão para o número máximo de iterações e inicia com k centróides aleatórios.

```
km <- function(dados,grupos){
  k <- kmeans(dados,grupos)
  print(table(iris$Species, k$cluster))
  plot(dados, col=k$cluster)
}
km(iris2,2)
```

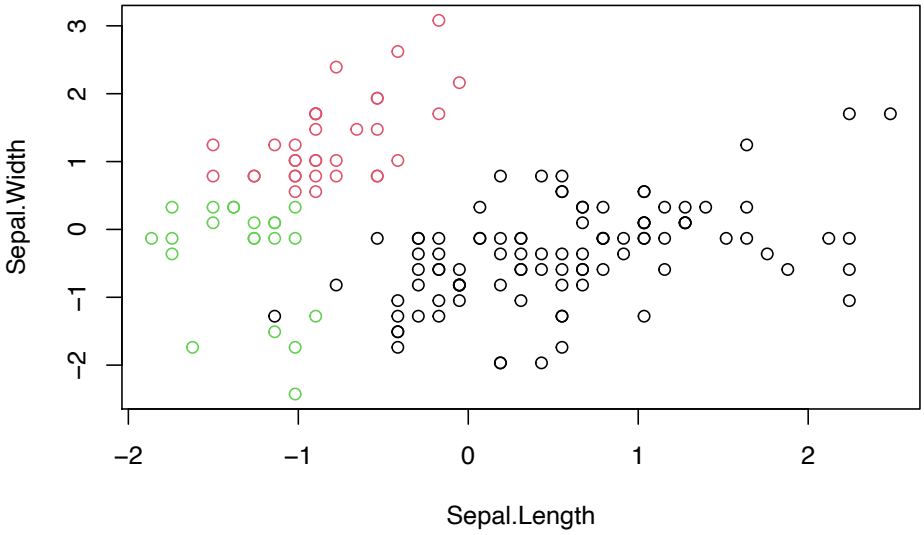
```
##
##                1  2
##   setosa      50  0
##   versicolor   0 50
##   virginica    0 50
```




```
km(iris2,3)
```

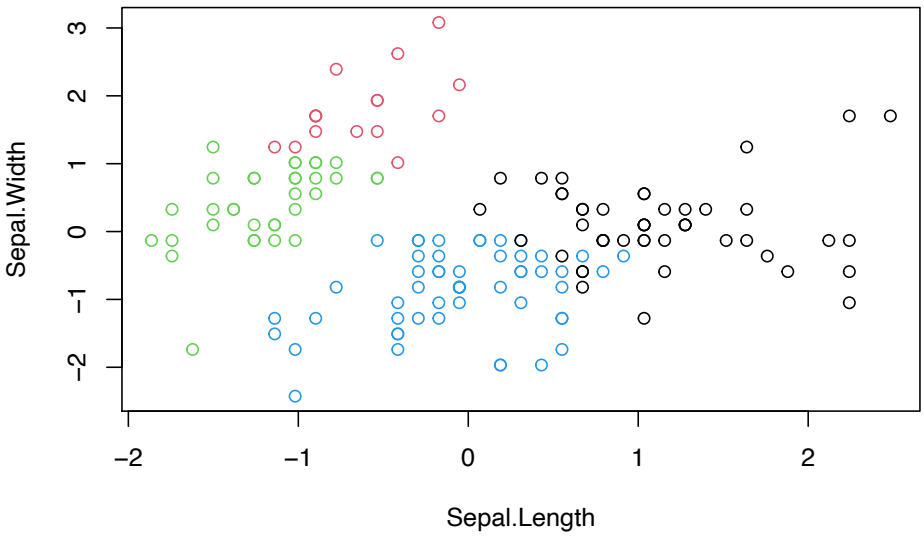
```
##
##                1  2  3
##   setosa       0 33 17
##   versicolor  46  0  4
```

```
##   virginica  50  0  0
```




```
km(iris2,4)
```

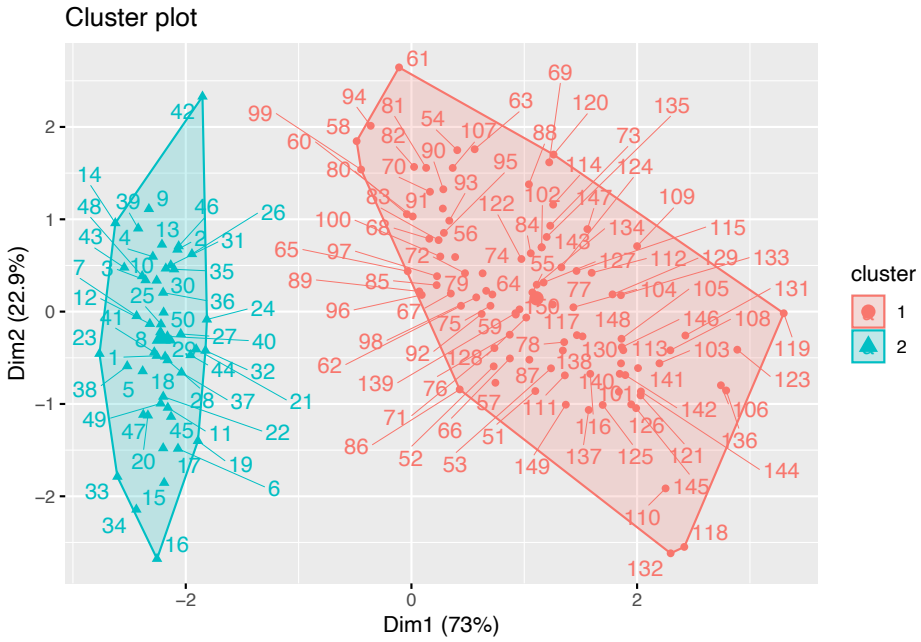
```
##
##                1  2  3  4
##   setosa       0 16 34  0
##   versicolor  11  0  0 39
##   virginica   36  0  0 14
```



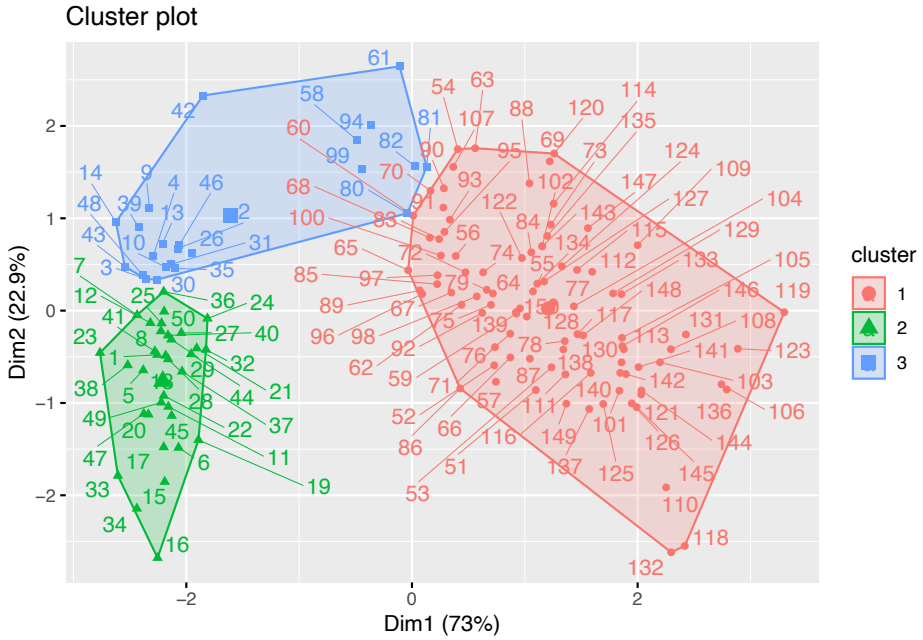
O pacote `factoextra` fornece uma série de melhorias para a análise de k-means. Além de gráficos mais sofisticados utilizando `ggplot2`, associa métodos

hierárquicos e métodos de particionamento.

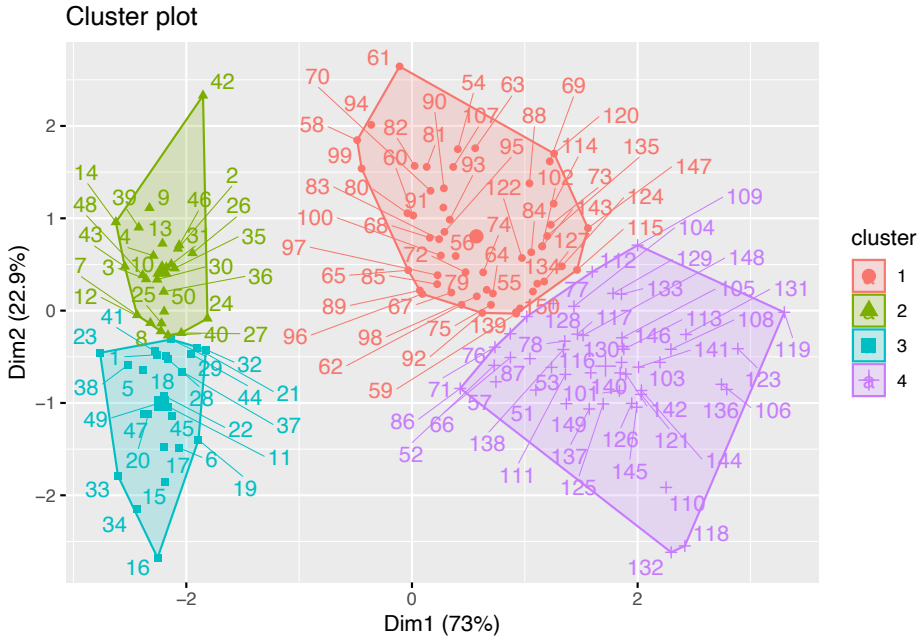
```
km2 <- function(dados,grupos){
  k <- kmeans(dados,grupos)
  fviz_cluster(k, iris2, repel = T)
}
km2(iris2,2)
```




```
km2(iris2,3)
```


Cluster plot
Dim2 (22.9%)
Dim1 (73%)
cluster
1
2
3

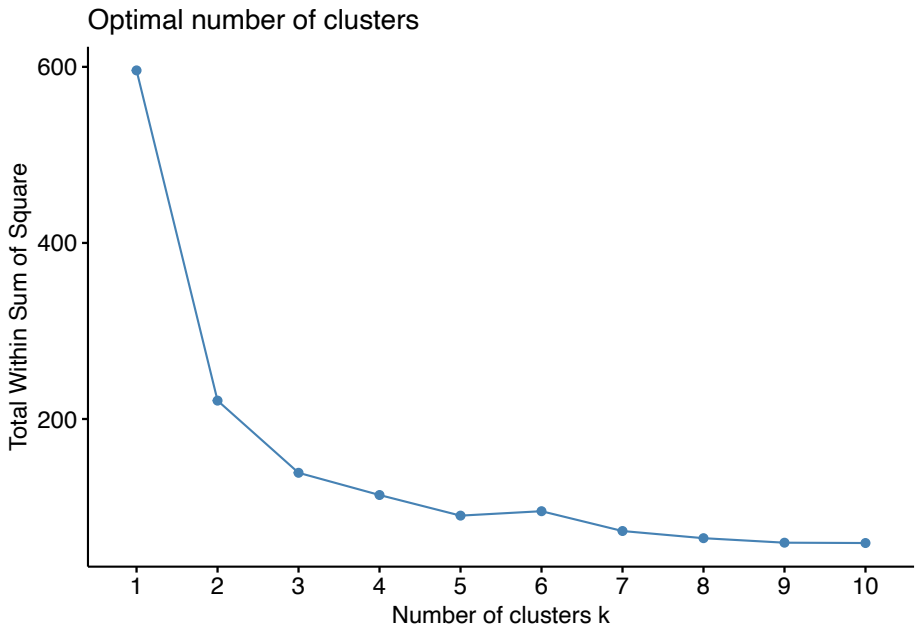
```
km2(iris2,4)
```


Cluster plot
Dim2 (22.9%)
Dim1 (73%)
cluster
1
2
3
4

Número de grupos

A função `factoextra::fviz_nbclust` fornece métodos para a escolha de um número ótimo de grupos. O método `wss` (*total* ***w****ithin* ***s****um of* ***s****quare*), busca um número de grupos que traga um bom custo-benefício entre o número de grupos (k) e a soma de quadrados total (SQT). Este custo-benefício é indicado onde a curva muda sua declividade, ou no 'cotovelo' (*elbow*) do gráfico de k por SQT. Tem suas origens no trabalho de (Thorndike, 1953).

```
fviz_nbclust(iris2, kmeans, method = 'wss')
```

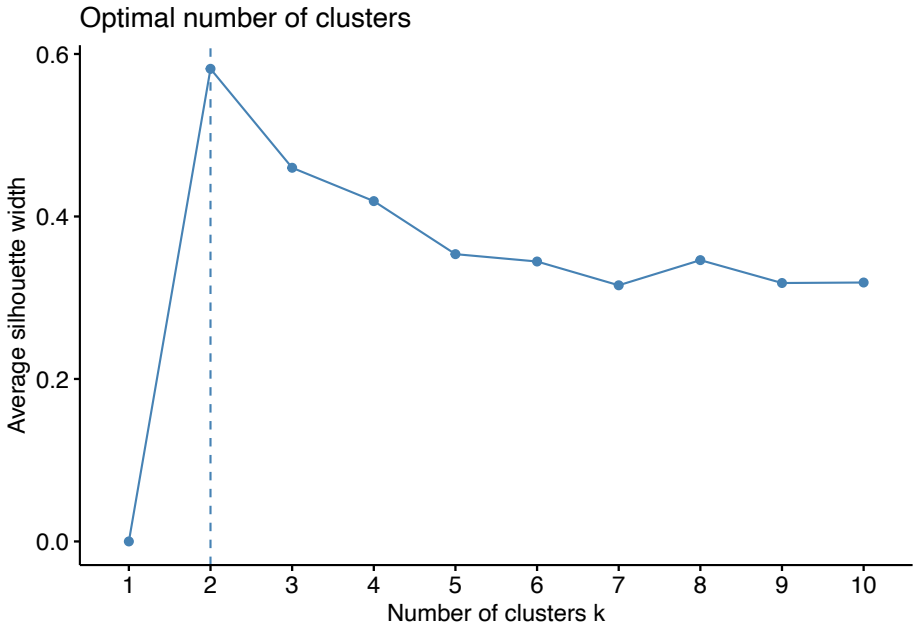


O método `silhouette` busca o número de grupos que maximize o tamanho médio da silhueta. É baseado em uma medida sugerida por (Rousseeuw and Kaufman, 1990), dada por

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{8.10}$$

- $-1 \leq s(i) \leq 1$
- $a(i)$: dissimilaridade média do elemento i em relação a todos os demais elementos do seu grupo A
- $d(i, C)$: dissimilaridade média do elemento i em relação a todos os elementos do grupo $C \neq A$
- $b(i) = \min_{C \neq A} d(i, C)$

```
fviz_nbclust(iris2, kmeans, method = 'silhouette')
```



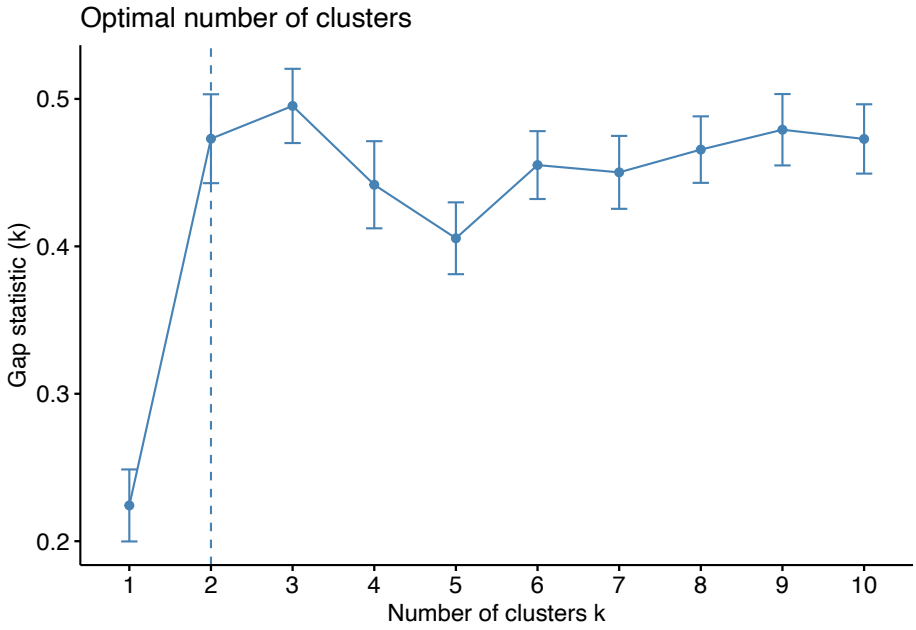
Proposto por (Tibshirani et al., 2001), o método `gap_stat` (*gap statistic*) compara variação total intra-grupo para diferentes valores de k com seus valores esperados sob alguma distribuição de referência. A estimativa dos clusters ótimos será o valor que maximiza a estatística de gap (isto é, que gera a maior estatística de gap).

$$\text{Gap}_n(k) = E^*\{\log(W_k)\} - \log(W_k) \tag{8.11}$$

- E^* é o valor esperado sob uma amostra de tamanho n da distribuição de referência
- $W_k = \sum_{j=1}^{k} \frac{1}{2n_j} D_j$
- $D_j = \sum_{i,i' \in C_j} d_{ii'}$
- $n_j = |C_j|$

```
fviz_nbclust(iris2, kmeans, method = 'gap_stat')
```

```
## Clustering k = 1,2,..., K.max (= 10): .. done
## Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
## .................................................. 50
## .................................................. 100
```



Exercício 8.7. Leia:
(a) a documentação da função `fviz_nbclust`.
(b) https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods.
(c) A Seção *e. Graphical Output Concerning Each Clustering*, pg. 83-88 de (Rousseeuw and Kaufman, 1990).

Exercício 8.8. Considere novamente o conjunto de dados `pib`.
(a) Verifique as sugestões do número ótimo de grupos com os diferente métodos disponíveis na função `fviz_nbclust`.
(b) Crie o grupamento que considerar mais adequado aos dados e apresente com a função `fviz_cluster`.
(c) Compare os resultados com o Exercício 7.3.

Exercício 8.9. Considere o conjunto de dados `drinks`, discutido no Capítulo 5.
(a) Calcule as distâncias de Manhattan, euclidiana e de Minkowski com $p = 3$.
(b) Obtenha os modelos hierárquicos utilizando as três distâncias do item (a). Você nota alguma diferença?
(c) Obtenha a seleção inicial dos centróides a partir de proposta de (Hartigan, 1975) apresentada na Equação (8.9). Sugestão: escreva uma função que dependa dos dados e de k, realizando alguma correção que considerar relevante.
(d) Calcule os centróides dos grupos obtidos no item (c).
(e) Calcule a VQI dos grupos obtidos no item (c) a partir da Eq. (8.7).
(f) Calcule a SQT dos grupos obtidos no item (c) a partir da Eq. (8.8).

(g) Verifique as sugestões do número ótimo de grupos com os diferente métodos disponíveis na função `fviz_nbclust`.
(h) Crie o grupamento que considerar mais adequado aos dados e apresente com a função `fviz_cluster`.

```
dat <- read.table('http://www.filipezabala.com/data/drinks.txt', header = T, sep = '\t')
```

Exercício 8.10. Considere o banco de dados sobre eficiência energética didcutido no Exemplo 5.2.
(a) Calcule as distâncias de Manhattan, euclidiana e de Minkowski com $p = 3$.
(b) Obtenha os modelos hierárquicos utilizando as três distâncias do item (a). (PODE DEMORAR!)
(c) Verifique as sugestões do número ótimo de grupos com os diferente métodos disponíveis na função `fviz_nbclust`.
(d) Crie o grupamento que considerar mais adequado aos dados e apresente com a função `fviz_cluster`.

```
library(readxl)
url1 <- 'http://archive.ics.uci.edu/ml/machine-learning-databases/00242/ENB2012_data.xlsx'
download.file(url1, 'temp.xlsx', mode = 'wb')
dat <- read_excel('temp.xlsx')
```

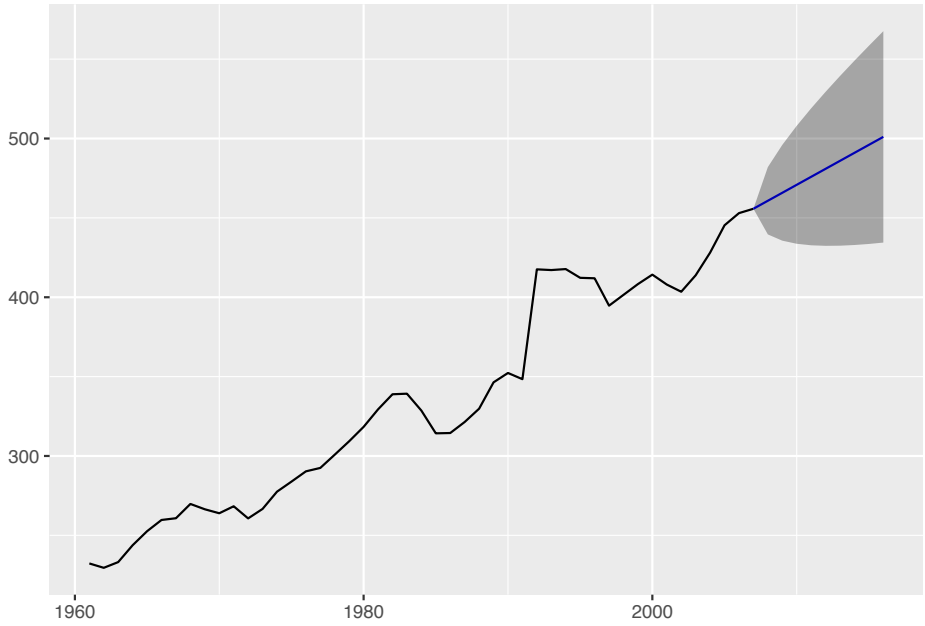
Chapter 9

Séries Temporais

Uma *série temporal* é um conjunto de dados observado no tempo. (Hyndman and Athanasopoulos, 2018) trazem uma compilação dos principais e mais recentes métodos da literatura, programados na biblioteca `fpp2` de (Hyndman, 2020). A seguir está o código `fits`, escrito com algumas funções desta biblioteca, de maneira a utilizar os seguintes métodos:

- ARIMA (*AutoRegressive Integrated Moving Average*), função `auto.arima`
- ETS (Modelo de espaço de estados com suavização exponencial), função `ets`
- TBATS (Modelo de espaço de estados com suavização exponencial com transformação Box-Cox, erros ARMA e componentes sazonais e de tendência), função `tbats`
- NNETAR (Rede neural autorregressiva), função `nnetar`

```
# calling package
library(jurimetrics)
# example
fits(livestock)
```




```
## $fcast
##      Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
## 2008            461   440   482   428   493
## 2009            466   436   496   420   512
## 2010            471   434   508   414   528
## 2011            476   433   519   410   542
## 2012            481   432   529   407   555
## 2013            486   432   539   404   568
## 2014            491   433   549   402   580
## 2015            496   434   558   401   591
## 2016            501   434   568   399   603
##
## $mse.pred
##   mse.pred.aa mse.pred.ets mse.pred.tb mse.pred.nn
## 1         109          103         253         742
##
## $best.model
## [1] "ets"
##
## $runtime
## Time difference of 2.34 secs
```

Exercício 9.1. Considere a função `fits`.
a) Avalie os parâmetros da função a partir do código.
b) Aplique nos bancos de dados `h02` e `gas`, utliizando os parâmetros que considerar mais adequados.

Exercício 9.2. Considere o banco de dados do índice Dow Jones disponível em https://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index, lido com o código abaixo.

```
url1 <- 'https://archive.ics.uci.edu/ml/machine-learning-databases/00312/dow_jones_index.zip'
download.file(url1, 'temp.zip', mode = 'wb')
dat <- suppressMessages(read_csv('temp.zip'))
st <- ts(dat$percent_change_next_weeks_price)
```

a. Modele e projete a série `percent_change_next_weeks_price` através da função `fits`. Busque pelos melhores modelos alterando os parâmetros da função, tais como `train` e `max.points`.

b. Avalie outras colunas do banco de dados e repita as operações do item a.

9.1 Impacto Causal

- http://google.github.io/CausalImpact/CausalImpact.html
- https://research.google/pubs/pub41854/
- https://github.com/klarsen1/MarketMatching

Chapter 10

Referências

Bibliography

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, (53):370–418.

Beasley, T. M. and Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *The Journal of Experimental Education*, 64(1):79–93.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2nd edition.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.

Bishop, C. M. (1999). Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 382–388.

Bolfarine, H. and Bussab, W. d. O. (2005). *Elementos de Amostragem*. Editora Blucher.

Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. R package version 1.3-0.

Chow, S.-C., Wang, H., and Shao, J. (2007). *Sample Size Calculations in Clinical Rsesearch, Second Edition*. CRC press.

de Finetti, B. (1974). Theory of Probability. a critical introductory treatment,(translation by a. machi and afm smith of 1970 book) 2 volumes.

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and statistics*. Pearson Education.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Ebbert, D. (2019). *chisq.posthoc.test: A Post Hoc Analysis for Pearson's Chi-Squared Test for Count Data*. R package version 0.1.2.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh and London.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science*, 1(2):115–137.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769.

Franz Faul, Edgard Erdfelder, A. B. and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41, 1149-1160*.

Franz Faul, Edgard Erdfelder, A.-G. L. and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39, 175-191*.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476.

Gannon, M. A., Pereira, C. A. d. B., and Polpo, A. (2019). Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels. *The American Statistician*, 73(sup1):213–222.

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hartl, D. L. and Fairbanks, D. J. (2007). Mud sticks: on the alleged falsification of mendel's data. *Genetics*, 175(3):975–979.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hyndman, R. (2020). *fpp2: Data for "Forecasting: Principles and Practice" (2nd Edition)*. R package version 2.4.

Hyndman, R. J. (1995). The problem with sturges rule for constructing histograms.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Izbicki, R. and dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*.

James, B. (2010). Probabilidade: Um Curso em Nível Intermediário, coleção euclides. *Rio de Janeiro. IMPA, 3a. Edição*.

Jessen, R. J. (1978). *Statistical Survey Techniques*.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the american statistical association*, 90(430):773–795.

Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions*. World Scientific.

Laplace, P.-S. (1825). Essai philosophique sur les probabilités (1814). *Printed as a preface to Théorie analytique des probabilités in the Oeuvres Complètes edition*.

Lindley, D. V. and Phillips, L. (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, 30(3):112–119.

Lloyd, S. P. (1957). Least squares quantization in PCM. *Tecnical note at Bell Laboratories in 1957, published after in IEEE transactions on information theory in 1982*, 28(2):129–137.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition.

Milošević, B. and Obradović, M. (2018). Comparison of efficiencies of some symmetry tests around an unknown centre. *Statistics*, 53(1):43–57.

Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (2005). *Applied Linear Statistical Models*. McGraw Hill/Irwin New York, 5 edition.

Novitski, E. (2004). On fisher's criticism of mendel's results with the garden pea. *Genetics*, 166(3):1133–1136.

Paula, G. A. (2013). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.

Paulino, C. D. M., Turkman, M. A. A., and Murteira, B. (2003). *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.

Pereira, C. A. d. B. and Stern, J. M. (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy*, 1(4):99–110.

Pereira, C. A. d. B. and Stern, J. M. (2020). The e-value: a fully Bayesian significance measure for precise statistical hypotheses and its research program.

Pereira, C. A. d. B., Stern, J. M., Wechsler, S., et al. (2008). Can a significance test be genuinely Bayesian? *Bayesian Analysis*, 3(1):79–100.

Pereira, C. A. d. B. and Wechsler, S. (1993). On the Concept of P-value. *Brazilian Journal of Probability and Statistics*, pages 159–177.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications, 2nd. edition*. John Wiley & Sons.

Rousseeuw, P. J. and Kaufman, L. (1990). Finding groups in data. *Hoboken: Wiley Online Library*.

Savage, L. J., Barnard, G., Cornfield, J., Bross, I., Good, I., Lindley, D., Cluniess-Ross, C., Pratt, J. W., Levene, H., Goldman, T., et al. (1962). On the foundations of statistical inference: Discussion. *Journal of the American Statistical Association*, 57(298):307–326.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Sheskin, D. J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures, 5th ed.* Chapman & Hall/CRC, Boca Raton, FL.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, pages 33–40.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567.

Tufte, E. R. (1993). Envisioning Information.

Tufte, E. R. (2006). *Beautiful Evidence*. Graphis Pr.

Tufte, E. R. and Graves-Morris, P. R. (1983). *The Visual Display of Quantitative Information*, volume 2. Graphics press Cheshire, CT.

Tufte, E. R., McKay, S. R., Christian, W., and Matey, J. R. (1998). Visual Explanations: Images and quantities, evidence and narrative.

Venables, W. N., Smith, D. M., Team, R. D. C., et al. (2020). An introduction to r.

Wechsler, S., Pereira, C. A. d. B., and Marques, P. C. F. (2008). Birnbaum's Theorem Redux.