

Jeidsan A. da C. Pereira

Estatística e Ciência de Dados

Notas e solução dos exercícios



Conteúdo

Prefácio	ix
Prefácio	ix
Notas sobre a edição	ix
Pendências	x
1 Estatística, Ciência de Dados e Megadados	1
1.1 Introdução	1
1.2 Aprendizado com estatística	1
1.3 Aprendizado automático	2
1.4 Uma cronologia do desenvolvimento da estatística	3
1.5 Notação e tipos de dados	3
1.6 Paradigmas para o aprendizado com estatística	3
1.7 Este livro	4
1.8 Conjuntos de dados	4
1.9 Notas do capítulo	4
I Análise Exploratória de Dados	5
2 Preparação dos dados	7
2.1 Considerações preliminares	7
2.2 Planilhas de dados	7
2.3 Construção de tabelas	7
2.4 Construção de gráficos	8
2.5 Notas de capítulo	8
2.6 Exercícios	8

3	Análise de dados de uma variável	21
3.1	Introdução	21
3.2	Distribuição de frequências	21
3.3	Medidas resumo	22
3.4	<i>Boxplots</i>	23
3.5	Modelos probabilísticos	24
3.6	Dados amostrais	24
3.7	Gráficos QQ	24
3.8	Desvio padrão e erro padrão	24
3.9	Intervalo de confiança e tamanho da amostra	25
3.10	Transformação de variáveis	25
3.11	Notas de capítulo	27
3.12	Exercícios	27
4	Análise de dados de duas variáveis	99
4.1	Introdução	99
4.2	Duas variáveis qualitativas	99
4.3	Duas variáveis quantitativas	99
4.4	Uma variável qualitativa e outra quantitativa	99
4.5	Notas de capítulo	99
4.6	Exercícios	99
5	Análise de dados de várias variáveis	101
5.1	Introdução	101
5.2	Gráficos para três variáveis	101
5.3	Gráficos para quatro ou mais variáveis	101
5.4	Medidas resumo multivariadas	101
5.5	Tabelas de contingência de múltiplas entradas	101
5.6	Notas de capítulo	101
5.7	Exercícios	101

6	Análise de Regressão	103
6.1	Introdução	103
6.2	Regressão linear simples	103
6.3	Regressão linear múltipla	103
6.4	Regressão para dados longitudinais	103
6.5	Regressão logística	103
6.6	Notas de capítulo	103
6.7	Exercícios	103
7	Análise de Sobrevivência	105
7.1	Introdução	105
7.2	Estimação da função de sobrevivência	105
7.3	Comparação de curvas de sobrevivência	105
7.4	Regressão para dados de sobrevivência	105
7.5	Notas de capítulo	105
7.6	Exercícios	105
II	Aprendizado Supervisionado	107
8	Regularização e Modelos Aditivos Generalizados	109
8.1	Introdução	109
8.2	Regularização	109
8.3	Modelos aditivos generalizados (GAM)	109
8.4	Notas de capítulo	109
8.5	Exercícios	109
9	Classificação por meio de técnicas clássicas	111
9.1	Introdução	111
9.2	Classificação por regressão logística	111
9.3	Análise discriminante linear	111
9.4	Classificador do vizinho mais próximo	111
9.5	Algumas extensões	111
9.6	Notas de capítulo	111
9.7	Exercícios	111

10 Algoritmos de Suporte Vetorial	113
10.1 Introdução	113
10.2 Fundamentação dos algoritmos de suporte vetorial	113
10.3 Classificador de margem máxima	113
10.4 Classificador de margem flexível	113
10.5 Classificador de margem não linear	113
10.6 Regressão por algoritmos de suporte vetorial	113
10.7 Notas de capítulo	113
10.8 Exercícios	113
11 Árvores e Florestas	115
11.1 Introdução	115
11.2 Classificação por árvores	115
11.3 <i>Bagging, boosting</i> e florestas	115
11.4 Árvores para regressão	115
11.5 Notas de capítulo	115
11.6 Exercícios	115
12 Redes neurais	117
12.1 Introdução	117
12.2 <i>Perceptron</i>	117
12.3 Redes com camadas ocultas	117
12.4 O algoritmo de retropropagação (<i>backpropagation</i>)	117
12.5 Aprendizado profundo (<i>Deep learning</i>)	117
12.6 Notas de capítulo	117
12.7 Exercícios	117
III Aprendizado não Supervisionado	119

13	Análise de Agrupamentos	121
13.1	Introdução	121
13.2	Estratégias de agrupamento	121
13.3	Algoritmos hierárquicos	121
13.4	Algoritmos de partição: K-médias	121
13.5	Notas de capítulo	121
13.6	Exercícios	121
14	Redução de dimensionalidade	123
14.1	Introdução	123
14.2	Análise de Componentes Principais	123
14.3	Análise fatorial	123
14.4	Análise de componentes independentes	123
14.5	Notas de capítulo	123
14.6	Exercícios	123
	Apêndice	123
A	Otimização numérica	125
A.1	Introdução	125
A.2	O método de Newton-Raphson	125
A.3	O método scoring	125
A.4	O método de Gauss-Newton	125
A.5	Métodos Quase-Newton	125
A.6	Aspectos computacionais	125
A.7	Notas de capítulo	125
A.8	Exercícios	125
B	Noções de simulação	127
B.1	Introdução	127
B.2	Método Monte Carlo	127
B.3	Simulação de variáveis discretas	127

B.4	Simulação de variáveis contínuas	127
B.5	Simulação de vetores aleatórios	127
B.6	Métodos de reamostragem	127
B.7	Notas de capítulo	127
B.8	Exercícios	127
C	Algoritmos para dados aumentados	129
C.1	Introdução	129
C.2	O algoritmo EM	129
C.3	O algoritmo EM Monte Carlo	129
C.4	Cálculo de erros padrões	129
C.5	O algoritmo para dados aumentados	129
C.6	Exercícios	129

Prefácio

Esta página contém notas e solução para os exercícios propostos no livro **Estatística e Ciência de Dados**, de autoria de Pedro Alberto Morettin e Júlio da Motta Singer, publicado pela LTC em 2022 [Morettin and Singer, 2022].

É importante destacar que trata-se de um produto não oficial, as anotações e soluções de exercícios aqui apresentadas são de cunho pessoal e não possuem qualquer revisão ou análise por parte dos autores da obra ou da editora. Dessa forma e por se tratar de um produto construído durante o processo de aprendizagem, o conteúdo pode conter erros, tanto no texto em si, como na lógica utilizada para solução dos exercícios.

Dúvidas ou sugestões de melhoria podem ser encaminhadas para o e-mail *jeidsan.pereira@gmail.com*¹.

Notas sobre a edição

- As notas constantes neste material são, em sua maioria, citações oriundas do livro em questão ou de livros e artigos relacionados. As citações indiretas são escritas em texto padrão, sem nenhum destaque, e são seguidas pela referência ao material consultado. Já as citações diretas, são destacadas no texto conforme a o exemplo abaixo.

Dado um número $0 < \alpha < 1$, a **média aparada** de ordem α , $\bar{x}(\alpha)$, é definida como a média do conjunto de dados obtido após a eliminação das 100% primeiras observações ordenadas e das 100% últimas observações ordenadas do conjunto original [...]. Para $\alpha = 0,25$ obtemos a chamada **meia média**. [Morettin and Singer, 2022, p. 47].

¹<mailto:jeidsan.pereira@gmail.com>

- Para esta edição, a numeração dos exercícios foi levemente alterada para fazer menção ao número do capítulo. Desta forma o exercício aqui numerado como Exercício 3.12, corresponde ao exercício de número 12 do capítulo de número 3. Também adicionamos a extensão `.xls` ou `.doc` aos conjuntos de dados disponibilizados pelo autor, facilitando a busca pelos arquivos no diretório de dados.
- Para facilitar a navegação, quando possível, adicionamos links para outras partes do texto que estejam relacionadas.
- Sempre que possível, procuramos também adicionar as referências originais aos conjuntos de dados, com o intuito de facilitar a consulta pelo leitor. Por exemplo, o conjunto de dados `rehabcardio.xls` é oriundo da pesquisa realizada por Carvalho et al. [2007].

Pendências

- Exercício 2.2;
- Revisar item b do Exercício 3.2;
-

1

Estatística, Ciência de Dados e Megadados

1.1 Introdução

Atualmente, os termos *Data Science* (**Ciência de Dados**) e *Big Data* (**Megadados**) são utilizados em profusão como se envolvessem conceitos novos, distintos daqueles com que os estatísticos lidam há cerca de dois séculos [Morettin and Singer, 2022, p. 1].

1.2 Aprendizado com estatística

O **aprendizado supervisionado** está relacionado com metodologia desenvolvida essencialmente para **previsão** e **classificação**. No âmbito da previsão, o objetivo é utilizar **variáveis preditivas** (sexo, classe social, renda, por exemplo) observadas em várias **unidades** (clientes de um banco, por exemplo) para “advinhar” valores de uma **variável resposta numérica** (saldo médio, por exemplo) de novas unidades. O problema de classificação consiste em qual categoria de uma **variável resposta qualitativa** (bons e maus pagadores, por exemplo) as novas unidades são classificadas [Morettin and Singer, 2022, p. 3].

No **aprendizado não supervisionado**, dispomos apenas um conjunto de dados, sem distinção entre preditoras e respostas, e o objetivo é descrever **associações** e **padrões** entre essas variáveis e **agrupá-las** com o objetivo de identificar características comuns e conjuntos de unidades de investigação ou desenvolver métodos para combiná-las e assim **reduzir sua dimensionalidade** [Morettin and Singer, 2022, p. 3].

Além de aprendizado supervisionado e não supervisionado, podemos acrescentar um terceiro tipo, denominado **aprendizado com reforço** (*reinforcement learning*), segundo o qual um algoritmo “aprende” a realizar determinadas tarefas por meio de repetições com o fim de maximizar um prêmio sujeito a um valor máximo [Morettin and Singer, 2022, p. 3].

Embora tanto o aprendizado supervisionado quanto o aprendizado com reforço utilizem um mapeamento entre entradas (*inputs*) e saídas (*outputs*), no primeiro caso a retroalimentação (*feedback*) fornecida ao algoritmo é um conjunto de ações corretas necessárias para a realização de uma tarefa; no aprendizado com reforço, por outro lado, a retroalimentação é baseada num sistema com prêmios e punições como indicativos de ações corretas ou incorretas [Morettin and Singer, 2022, p. 3].

1.3 Aprendizado automático

Jordan [2019 *apud* Morettin and Singer, 2022, p. 4] distingue três tipos de inteligência artificial: i) inteligência artificial imitativa humana; ii) aumento de inteligência; e iii) infraestrutura inteligente.

De modo informal, a inteligência artificial está relacionada com um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos (Chollet, 2018) e consequentemente, intimamente ligada ao desenvolvimento da computação (ou programação de computadores) [Morettin and Singer, 2022, p. 4].

Convém ressaltar que o objetivo do aprendizado automático não é o mesmo daquele considerado na análise de regressão usual, em que se pretende entender como cada variável preditora X_0 está associada com a variável resposta. O objetivo do aprendizado automático é selecionar o modelo que produz melhores previsões, mesmo que as variáveis selecionadas com essa finalidade não sejam aquelas consideradas numa análise padrão [Morettin and Singer, 2022, p. 5].

1.4 Uma cronologia do desenvolvimento da estatística

Sem notas para esta seção.

1.5 Notação e tipos de dados

Sem notas para esta seção.

1.6 Paradigmas para o aprendizado com estatística

Sem notas para esta seção.

1.7 Este livro

Independentemente do volume de dados disponíveis para análise, Ciência de Dados é uma atividade multidisciplinar que envolve: i) um problema a ser resolvido com questões claramente especificadas; ii) um conjunto de dados (seja ele volumoso ou não); iii) os meios para sua obtenção; iv) sua organização; v) a especificação do problema original em termos das variáveis desse conjunto de dados; vi) a descrição e resumo dos dados à luz do problema a ser resolvido; vii) a escolha das técnicas estatísticas apropriadas para a resolução desse problema; viii) os algoritmos computacionais necessários para a implementação dessas técnicas; ix) a apresentação dos resultados [Moret-tin and Singer, 2022, p. 11].

1.8 Conjuntos de dados

Sem notas para esta seção.

1.9 Notas do capítulo

Sem notas para esta seção.

Parte I

Análise Exploratória de Dados



2

Preparação dos dados

2.1 Considerações preliminares

O ramo da Estatística conhecido como **Análise Exploratória de Dados** se ocupa da organização e resumo dos dados de uma amostra ou, eventualmente, de toda a população e o ramo conhecido como **Inferência Estatística** se refere ao processo de se tirar conclusões sobre uma população com base em uma amostra dela [Morettin and Singer, 2022, p. 21].

2.2 Planilhas de dados

Sem notas para esta seção.

2.3 Contrução de tabelas

Sem notas para esta seção.

2.4 Construção de gráficos

Sem notas para esta seção.

2.5 Notas de capítulo

Sem notas para esta seção.

2.6 Exercícios

Exercício 2.1

O objetivo de um estudo da Faculdade de Medicina da USP foi avaliar a associação entre a quantidade de morfina administrada a pacientes com dores intensas provenientes de lesões medulares ou radiculares e a dosagem dessa substância em seus cabelos. Três medidas foram realizadas em cada paciente, a primeira logo após o início do tratamento e as demais após 30 e 60 dias. Detalhes podem ser obtidos no documento disponível no arquivo `morfina.doc`.

A planilha `morfina.xls`, disponível no arquivo `morfina` foi entregue ao estatístico para análise e contém resumos de características demográficas além dos dados do estudo.

- Com base nessa planilha, apresente um dicionário com a especificação das variáveis segundo as indicações da Seção 2.2 e construa a planilha correspondente.
- Com as informações disponíveis, construa tabelas para as variáveis sexo, raça, grau de instrução e tipo de lesão segundo as sugestões da Seção 2.3.

Solução. Utilizando o arquivo `morfina.doc` chegamos à seguinte solução para o item a:

Tabela 2.1: Dicionários para as variáveis do estudo *morfina.doc*¹

Rótulo	Variável	Unidade de medida
id	Identificação do paciente	
data	Data de avaliação do paciente	
idade	Idade do paciente	anos
sexo	Sexo do paciente	1 - masculino 2 - feminino
raça	Raça e/ou etnia do paciente	1 - pardo 2 - negro 3 - branco 4 - indígena 5 - amarelo
religião	Religião do paciente	1 - catolico 2 - sem religião 3 - evangélico 4 - espírita 5 - judeu 6 - outra
peso	Peso do paciente	quilogramas (kg)
altura	Altura do paciente	metros (m)
instrução	Grau de instrução do paciente	1 - Analfabeto 2 - Alfabetizado 3 - Ens. Fundamental 4 - Ens. Médio 5 - Ens. Superior
enfermidade	Enfermidade primária do paciente	1 - Lesão medular 2 - Lesão radicular 3 - Trauma 4 - FAF 5 - Pós cirúrgico
tipo	Tipo de cabelo do paciente	1 - Natural 2 - Artificial
cor	Cor do cabelo do paciente	1 - Marrom 2 - Vermelho 3 - Preto 4 - Louro 5 - Cinza 6 - Branco 7 - Outro
forma	Forma do cabelo do paciente	1 - Caucasiana 2 - Asiática 3 - Negróide 4 - Outra
medicamentos	Medicamentos em uso pelo paciente	
composicao	Composição da solução do reservatório	
ampola	Número da ampola de morfina	

Rótulo	Variável	Unidade de medida
sf	Quantidade de soro fisiológico para diluição	mililitros (ml)
concentracao	Concentração da solução	percentual (%)
disp_prop	Número de disparos proposto por dia	
disp_real	Número de disparos realizados por dia	
vol_desprez	Quantidade de solução desprezada no reservatório	mililitros (ml)
con_desp	Concentração da solução desprezada no reservatório	percentual (%)
obstrucao	Ocorrência de obstrução no cateter	1 - Sim 2 - Não
infeccao	Ocorrência de infecção do sistema	1 - Sim 2 - Não
nausea	Ocorrência de náusea como efeito colateral	1 - Sim 2 - Não
sonolencia	Ocorrência de sonolência como efeito colateral	1 - Sim 2 - Não
constipação	Ocorrência de constipação como efeito colateral	1 - Sim 2 - Não
tontura	Ocorrência de tontura como efeito colateral	1 - Sim 2 - Não
prurido	Ocorrência de prurido como efeito colateral	1 - Sim 2 - Não
retencao	Ocorrência de retenção urinária como efeito colateral	1 - Sim 2 - Não
outros	Ocorrência de outros efeitos colaterais	1 - Sim 2 - Não
d0	Dose inicial de morfina	miligrama (mg)
d30	Dose de morfina após 30 dias	miligrama (mg)
d60	Dose de morfina após 60 dias	miligrama (mg)
t0	Quantidade inicial de morfina no cabelo	miligrama (mg)
t30	Quantidade de morfina no cabelo após 30 dias	miligrama (mg)
t60	Quantidade de morfina no cabelo após 60 dias	miligrama (mg)

Para o item b, como teremos quatro tabelas distintas e os dados não estão cruzados, optamos por incluir o rótulo da primeira coluna no título do gráfico, a fim de não termos títulos repetidos. Temos o seguinte:

Tabela 2.2: Distribuição dos pacientes conforme o sexo

Sexo	Número de pacientes	Percentual (%)
Masculino	19	54
Feminino	16	46
Total	35	100

¹Trata-se de uma versão inicial para o dicionário de dados. O mesmo será revisado posteriormente para se adequar à especificação do documento *morfina.doc*.

Tabela 2.3: Distribuição dos pacientes conforme raça/etnia

Raça	Número de pacientes	Percentual (%)
Pardo	20	57
Branco	13	37
Negro	2	6
Total	35	100

Tabela 2.4: Distribuição dos pacientes conforme o grau de instrução

Grau de Instrução	Número de pacientes	Percentual (%)
Ensino Fundamental	25	71
Ensino Médio	9	26
Ensino Superior	1	3
Total	35	100

Tabela 2.5: Distribuição dos pacientes conforme o tipo de lesão

Tipo de lesão	Número de pacientes	Percentual (%)
Medular	10	29
Radicular	25	71
Total	35	100

Exercício 2.2

A Figura 2.6 foi extraída de um estudo sobre atitudes de profissionais de saúde com relação a cuidados com infecção hospitalar. Critique-a e reformule-a para facilitar sua leitura, lembrando que a comparação de maior interesse é entre as diferentes categorias profissionais.

Solução.

Tabela 2.6: ABC

A	B	C
A	B	C
A	B	C

Exercício 2.3

Utilize as sugestões para construção de planilhas apresentadas na Seção 2.2 com a finalidade de preparar os dados do arquivo empresa para análise estatística.

Solução. Vamos iniciar construindo um dicionário para os dados

Tabela 2.7: Dicionário de dados para a planilha `empresa.xls`

Rótulo	Descrição	Unidade de medida
id	Identificador ddo funcionário	
estado	Estado civil do funcionário	1 - Solteiro 2 - Casado
instrucao	Grau de instrução do funcionário	1 - Ensino Fundamental 2 - Ensino Médio 3 - Ensino Superior
filhos	Número de filhos do funcionário	
salario	Salário do funcionário	salário mínimo
anos	Idade do funcionário	anos
meses	Fração da idade do funcionário	meses
regiao	Região de procedência do funcionário	1 - Interior 2 - Capital 3 - Outra

Com o dicionário de dados em mãos, podemos atualizar a planilha:

Tabela 2.8: Dados de funcionários de uma empresa (parte)

id	estado	instrucao	filhos	salario	anos	meses	regiao
1	1	1		4.00	26	3	1
2	2	1	1	4.56	32	10	2
3	2	1	2	5.25	36	5	2
4	1	2		5.73	20	10	3
5	1	1		6.26	40	7	3
6	2	1	0	6.66	28	0	1
7	1	1		6.86	41	0	1
8	1	1		7.39	43	4	2
9	2	2	1	7.59	34	10	2
10	1	2		7.44	23	6	3

Exercício 2.4

Num estudo planejado para avaliar o consumo médio de combustível de veículos em diferentes velocidades foram utilizados 4 automóveis da marca A e 3 automóveis da marca B selecionados ao acaso das respectivas linhas de produção. O consumo (em

L/km) de cada um dos 7 automóveis foi observado em 3 velocidades diferentes (40 km/h, 80 km/h e 110km/h) Delineie uma planilha apropriada para a coleta e análise estatística dos dados, rotulando-a adequadamente.

Solução. Vamos começar construindo um dicionário de dados para a planilha:

Tabela 2.9: Dicionário para as variáveis do estudo sobre consumo de combustível

Rótulo	Descrição	Unidade
id	Identificador do veículo	
marca	Marca fabricante do veículo	
modelo	Modelo do veículo	
consumo40	Consumo de combustível a 40 km/h	L/km
consumo80	Consumo de combustível a 80 km/h	L/km
consumo110	Consumo de combustível a 110 km/h	L/km

Agora vamos montar uma planilha (fictícia) seguindo a padronização definida acima:

Tabela 2.10: Planilha de dados para estudo sobre consumo de combustível

id	marca	modelo	consumo40	consumo80	consumo110
1	A	XPTO 1			
2	A	XPTO 2			
3	A	XPTO 3			
4	A	XPTO 4			
5	B	XYZ 1			
6	B	XYZ 2			
7	B	XYZ 3			

Exercício 2.5

Utilizando os dados do arquivo `enforco.xls`, prepare uma planilha Excel num formato conveniente para análise pelo R. Inclua apenas as variáveis Idade, Altura, Peso, Frequência cardíaca e V_{O2} no repouso além do quociente VE/VC_{O2} , as correspondentes porcentagens relativamente ao máximo, o quociente V_{O2}/FC no pico do exercício e data do óbito. Importe a planilha Excel que você criou utilizando comandos R e obtenha as características do arquivo importado (número de casos, número de observações omissas etc.)

Solução. Conforme especificação enunciada, criamos o arquivo `esforco.csv` que poderá ser carregado da seguinte maneira:

```
(esforco <- read_csv(paste0(data_dir, "esforco.csv")))
```

```
## Rows: 127 Columns: 8
## -- Column specification -----
## Delimiter: ","
## chr (1): obito
## dbl (7): id, idade, altura, peso, fc_repouso, vo2_repouco, ve_vo2_pico
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 127 x 8
##       id idade altura peso fc_repouso vo2_repouco ve_vo2_pico obito
##   <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl> <chr>
## 1     1     38    149    54      89     5.9    65.6 26/07/1991
## 2     2     49    167    80      69     3.4    37.3 30/07/1995
## 3     3     65    153    56      82     3     59.7 21/08/1993
## 4     4     52    175    78      89     3.8    52.4 14/11/1992
## 5     5     52    157    59      82     3.2    48.8 30/07/1994
## 6     6     58    150    62      75     3.8    54.1 Não
## 7     7     24    155    42      89     3.5    102. 17/10/1991
## 8     8     39    149    55      91     3.9    67.8 31/08/1992
## 9     9     48    160    77     101     2.5    59.5 Não
## 10    10     50    171    81     120     3     47.8 Não
## # i 117 more rows
```

Temos 127 casos e 8 variáveis.

Exercício 2.6

A Figura 2.7 contém uma planilha encaminhada pelos investigadores responsáveis por um estudo sobre AIDS para análise estatística. Organize-a de forma a permitir sua análise por meio de um pacote computacional como o R.

Solução. Os dados foram reorganizados no arquivo `aids.csv`.

```
(aids <- read_csv(paste0(data_dir, "aids.csv")))
```

```
## Rows: 19 Columns: 7
```



```
## -- Column specification -----
##
## Delimiter: ","
## chr (3): id, dst, mac
## dbl (4): grupo, diagnostico, peso, tempo_peso
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

## # A tibble: 19 x 7
##   id      grupo diagnostico dst      mac      peso tempo_peso
##   <chr>    <dbl>    <dbl> <chr>    <chr>    <dbl>    <dbl>
## 1 2847111D      1      0 <NA>    pilula     11      37
## 2 3034048F      1     0.5 <NA>    pilula     NA      NA
## 3 3244701J      1      1 <NA>    condon     NA      NA
## 4 2943791B      1      0 <NA>    não        8      39
## 5 3000327F      1      4 condiloma/sífilis não        9      39
## 6 3232893D      1      1 <NA>    diu         3      39
## 7 3028772E      1      3 <NA>    não         3      38
## 8 3240047G      1      0 <NA>    pilula      9      38
## 9 3017222G      1     NA  HPV      condon     NA      NA
## 10 3015834J      1      2 <NA>    condon     14      40
## 11 3173611E      2     0.4 abcesso ovariano condon     15      40
## 12 3296159D      2      0 <NA>    condon     NA      NA
## 13 3147820D1      2      2 <NA>    <NA>        4      37
## 14 3274750K      2      3 <NA>    condon      8      38
## 15 3274447H      2      0 sífilis com 3 meses condon     NA      NA
## 16 2960066D      2      5 <NA>    <NA>       13      36
## 17 3235727J      2      7 <NA>    condon     -2      38
## 18 3264897E      2      0 condiloma/sífilis condon      0      NA
## 19 3044120J      2      5 HPV      <NA>        3      39
```

Uma possível melhoria seria a transformação das variáveis `grupo` e `mac` em fatores.

Exercício 2.7

A planilha apresentada na Figura 2.8 contém dados de um estudo em que o limiar auditivo foi avaliado nas orelhas direita (OD) e esquerda (OE) de 13 pacientes em 3 ocasiões (Limiar, Teste 1 e Teste 2). Reformate-a segundo as recomendações da Seção 2.2, indicando claramente

- a definição das variáveis.
- os rótulos para as colunas da planilha.

Solução. Precisamos inicialmente definir um dicionário para as variáveis e, na sequência, refatorar a planilha.

Tabela 2.11: Tabela 2.11: Dicionário de dados para o estudo sobre limiar auditivo ²

Rótulo	Descrição da variável	Unidade de medida
id	Identificador do paciente	
od0	Limiar auditivo da orelha direita no início do estudo	%
oe0	Limiar auditivo da orelha esquerda no início do estudo	%
od1	Limiar auditivo da orelha direita no primeiro teste	%
oe1	Limiar auditivo da orelha esquerda no primeiro teste	%
od2	Limiar auditivo da orelha direita no segundo teste	%
oe2	Limiar auditivo da orelha esquerda no segundo teste	%

Tabela 2.12: Tabela 2.12: Limiar auditivo de pacientes observados em 3 ocasiões

id	od0	oe0	od1	oe1	od2	oe2
1	50.00	50.00	50.00	50.00	80.00	80.00
2	41.00	40.00	45.00	50.00	68.00	80.00
3	41.25	41.25	45.00	45.00	64.00	72.00
4	45.00	43.75	60.00	50.00	76.00	88.00
5	51.25	47.50	50.00	50.00	80.00	80.00
6	45.00	52.50	50.00	50.00	84.00	96.00
7	52.50	50.00	55.00	45.00	40.00	28.00
8	42.15	48.75	50.00	50.00	80.00	76.00
9	50.00	48.75	50.00	50.00	72.00	80.00
10	47.50	46.25	55.00	60.00	84.00	84.00
11	55.00	56.25	40.00	35.00	80.00	84.00
12	46.25	46.25	45.00	45.00	72.00	84.00
13	50.00	47.50	40.00	50.00	76.00	76.00

Exercício 2.8

A planilha disponível no arquivo `idades.xls` contém informações demográficas de 3554 municípios brasileiros.

²Como consideramos o limite de detecção como sendo 0.05, foram utilizadas duas casas decimais para representar os limites auditivos observados.

- Importe-a para permitir a análise por meio do software R, indicando os problemas encontrados nesse processo além de sua solução.
- Use o comando `summary` para obter um resumo das variáveis do arquivo.
- Classifique cada variável como numérica ou alfanumérica e indique o número de observações omissas de cada uma delas.

Solução. Ao tentar realizar a leitura utilizando a função `read_csv` que já conhecemos, obteríamos um erro devido a planilha conter formatações.

```
idades <- read_csv(paste0(data_dir, "idades.xls"))
```

```
## Error in vroom_(file, delim = delim %||% col_types$delim, col_names = col_names, : cadeia de caracteres com nul inc
```

Podemos limpar a formatação da planilha e tentar a importação novamente ou utilizar a função `read_xls` do pacote **readxl**.

```
idades <- readxl::read_xls(paste0(data_dir, "idades.xls"), na = '-')
```

Note que, como os valores faltantes na planilha estão indicado com um hífen, utilizamos o argumento `na = '-'` para convertelos automaticamente para NA. Note também que as últimas duas linhas do data frame `idades` contém os totalizadores e não observações. Vamos removê-las!

```
idades <- head(idades, -2)
```

Note que o rótulo das variáveis está em maiúsculo, vamos colocá-los em minúsculo com a ajuda da função `str_to_lower()` do pacote **stringr**:

```
colnames(idades) <- str_to_lower(colnames(idades))
```

Podemos agora ver que temos 17 variáveis e 3554 unidades de análise. Um resumo das variáveis do conjunto de dados é mostrado pelo comando `summary`:

```
summary(idades)
```

```
##      munic      uf      código      poptot
## Length:3554   Length:3554   Min.   :1001   Min.    :    795
```

```
## Class :character   Class :character   1st Qu.:1889   1st Qu.:   7995
## Mode  :character   Mode  :character   Median :3720   Median :  15632
##                                     Mean  :3440   Mean   :  43650
##                                     3rd Qu.:4609   3rd Qu.:  30655
##                                     Max.   :5497   Max.    :10406166
##
##      cres_pop      popurb      pibtot      cres_pib
## Min.   :-13.330   Min.    :   423   Min.    :   0.90   Min.    : 0.0000
## 1st Qu.:  0.020   1st Qu.:  4388   1st Qu.:   13.48   1st Qu.: 0.6936
## Median :  1.145   Median :  9232   Median :   26.79   Median : 1.0372
## Mean   :  1.283   Mean    : 36908   Mean    :  177.93   Mean    : 1.1607
## 3rd Qu.:  2.310   3rd Qu.: 20732   3rd Qu.:   66.80   3rd Qu.: 1.4493
## Max.    : 23.630   Max.    :9785640   Max.    :105906.65   Max.    :24.6598
##                                     NA's    :14      NA's    :14
##      grau1      grau2      superior      lloumais
## Min.    :   469   Min.    :   47   Min.    :    0   Min.    :   37
## 1st Qu.:  4738   1st Qu.:  495   1st Qu.:   75   1st Qu.:  407
## Median :  8491   Median :  950   Median :   178   Median :   786
## Mean    : 22833   Mean    : 5060   Mean    :  2064   Mean    :  5407
## 3rd Qu.: 16057   3rd Qu.: 2272   3rd Qu.:   522   3rd Qu.:  1963
## Max.    :5322497   Max.    :1606381   Max.    :1076916   Max.    :2142313
## NA's    :13      NA's    :13      NA's    :13      NA's    :13
##      empregad      microemp      peqemp      medemp
## Min.    :   10   Min.    :   3.0   Min.    :   0.00   Min.    : 0.000
## 1st Qu.:  414   1st Qu.:  94.0   1st Qu.:   1.00   1st Qu.: 1.000
## Median :   926   Median : 207.0   Median :   3.00   Median : 1.000
## Mean    :  7778   Mean    : 916.9   Mean    :  36.69   Mean    :  6.929
## 3rd Qu.:  2743   3rd Qu.: 503.0   3rd Qu.:  13.00   3rd Qu.:  2.000
## Max.    :3986021   Max.    :377600.0   Max.    :18494.00   Max.    :3198.000
## NA's    :14      NA's    :14      NA's    :14      NA's    :14
##      graenp
## Min.    : 0.000
## 1st Qu.: 0.000
## Median : 0.000
## Mean    : 1.341
## 3rd Qu.: 1.000
## Max.    :568.000
## NA's    :14
```

Esse comando também nos permite perceber os tipos de cada uma das variáveis e se as mesmas contém valores faltantes. Essas informações estão resumidas na Tabela 2.13.

Tabela 2.13: Tabela 2.13: Resumo das observações da tabela `idades.xls`

Variável	Tipo	Número de observações faltantes
<code>munic</code>	Alfanumérica	0
<code>uf</code>	Alfanumérica	0
<code>codigo</code>	Numérica	0
<code>poptot</code>	Numérica	0
<code>cres_pop</code>	Numérica	0
<code>popurb</code>	Numérica	0
<code>pibtot</code>	Numérica	14
<code>cres_pib</code>	Numérica	14
<code>grau1</code>	Numérica	13
<code>grau2</code>	Numérica	13
<code>superior</code>	Numérica	13
<code>iloumais</code>	Numérica	13
<code>empregad</code>	Numérica	14
<code>microemp</code>	Numérica	14
<code>peqemp</code>	Numérica	14
<code>medemp</code>	Numérica	14
<code>graemp</code>	Numérica	14

Exercício 2.9

Preencha a ficha de inscrição do Centro de Estatística Aplicada www.ime.usp.br/-cea³ com as informações de um estudo em que você está envolvido.

Solução. Não se aplica.

³<http://www.ime.usp.br/-cea>



3

Análise de dados de uma variável

3.1 Introdução

A ideia de uma análise descritiva de dados é tentar responder as seguintes questões:

- i) Qual a frequência com que cada valor (ou intervalo de valores) aparece no conjunto de dados ou seja, qual a distribuição de frequências dos dados?
- ii) Quais são alguns valores típicos do conjunto de dados, como mínimo e máximo?
- iii) Qual seria um valor para representar a posição (ou localização) central do conjunto de dados?
- iv) Qual seria uma medida da variabilidade ou dispersão dos dados?
- v) Existem valores atípicos ou discrepantes (outliers) no conjunto de dados?
- vi) A distribuição de frequências dos dados pode ser considerada simétrica?

[Morettin and Singer, 2022, p. 37]

3.2 Distribuição de frequências

Sem notas para esta seção.

3.3 Medidas resumo

Dado um número $0 < \alpha < 1$, a **média aparada** de ordem α , $\bar{x}(\alpha)$, é definida como a média do conjunto de dados obtido após a eliminação das 100% primeiras observações ordenadas e das 100% últimas observações ordenadas do conjunto original [...]. Para $\alpha = 0,25$ obtemos a chamada **meia média**. [Morettin and Singer, 2022, p. 47].

Dado um número natural $k \geq 2$ e um conjunto $X = \{x_1, \dots, x_n\}$ com $n \in \mathbb{N}$, o k -ésimo **momento centrado** de X é dado por [Morettin and Singer, 2022, p. 50]:

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

Dentre as medidas de assimetria, as mais comuns são:

- a) o coeficiente de assimetria de Fisher-Pearson:

$$g_1 = \frac{m_3}{m_2^{(3/2)}}$$

- b) o coeficiente de assimetria de Fisher-Pearson ajustado:

$$\frac{\sqrt{n(n-1)}}{n-2} g_1$$

[Morettin and Singer, 2022, p. 50]

As principais propriedades desses coeficientes são

- i) seu sinal reflete a direção da assimetria (sinal negativo corresponde a assimetria à direita e sinal positivo corresponde a assimetria à esquerda);

- ii) comparam a assimetria dos dados com aquela da distribuição normal, que é simétrica,
- iii) valores mais afastados do zero indicam maiores magnitudes de assimetria e consequentemente, maior afastamento da distribuição normal;
- iv) a estatística indicada em (3.15) tem um ajuste para o tamanho amostral;
- v) esse ajuste tem pequeno impacto em grandes amostras. [Morettin and Singer, 2022, p. 51]

Outro coeficiente de assimetria mais intuitivo é o chamado **coeficiente de assimetria de Pearson 2**, estimado por

$$Sk_2 = \frac{3 \cdot [x - med(x_1, \dots, x_n)]}{S}$$

[Morettin and Singer, 2022, p. 51].

Seja X uma variável aleatória qualquer, com média μ e variância σ^2 . A **curtose** de X é definida por

$$K(X) = E \left[\frac{(X - \mu)^4}{\sigma^4} \right]$$

[Morettin and Singer, 2022, p. 53]

3.4 Boxplots

Sem notas para esta seção.

3.5 Modelos probabilísticos

Sem notas para esta seção.

3.6 Dados amostrais

Sem notas para esta seção.

3.7 Gráficos QQ

Uma das questões fundamentais na especificação de um modelo para inferência estatística é a escolha de um modelo probabilístico para representar a distribuição (desconhecida) da variável de interesse na população. Uma possível estratégia para isso é examinar o histograma dos dados amostrais e compará-lo com histogramas teóricos associados a modelos probabilísticos candidatos. Alternativamente, os gráficos QQ (QQ plots) também podem ser utilizados com essa finalidade [Morettin and Singer, 2022, p. 59].

3.8 Desvio padrão e erro padrão

Sem notas para esta seção.

3.9 Intervalo de confiança e tamanho da amostra

[...] **margem de erro**, que, essencialmente, é uma medida de nossa incerteza na extrapolação dos resultados obtidos para a população de onde assumimos que foi obtida [Morettin and Singer, 2022, p. 65].

A margem de erro depende do processo amostral, do desvio padrão amostral S , do tamanho da amostra n e é dado por $me = \frac{kS}{\sqrt{n}}$ em que k é uma constante que depende do modelo probabilístico adotado e da confiança com que pretendemos fazer a inferência [Morettin and Singer, 2022, p. 65].

Especificamente no caso da estimação da média (populacional) μ de uma variável X , a pergunta seria *Qual é o tamanho da amostra necessário para que a estimativa \bar{X} da média μ tenha uma precisão ε ?* A resposta pode ser obtida da expressão (3.21), fazendo $\varepsilon = \frac{1,96 \cdot S}{\sqrt{n}}$ [Morettin and Singer, 2022, p. 66].

3.10 Transformação de variáveis

Se quisermos utilizar os procedimentos talhados para análise de dados com distribuição normal em situações nas quais a distribuição

dos dados amostrais é sabidamente assimétrica, pode-se considerar uma transformação das observações com a finalidade de se obter uma distribuição “mais simétrica” e portanto, mais próxima da distribuição normal. Uma transformação bastante usada com esse propósito é

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \log(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0 \end{cases}$$

Essa transformação com $0 < p < 1$ apropriada para distribuições assimétricas à direita, pois valores grandes decrescem de x decrescem mais relativamente a valores pequenos. Para distribuições assimétricas à esquerda, basta tomar $p > 1$. Normalmente, consideramos valores de p na sequência

$$\dots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \dots$$

e para cada um deles construímos gráficos apropriados (histogramas, boxplots) com os dados originais transformados, com a finalidade de escolher o valor mais adequado para p . Hinkley (1977) sugere que para cada valor de p na sequência acima se calcule a média, a mediana e um estimador de escala (esvio padrão ou algum estimador robusto) e então se escolha o valor que minimiza

$$d_p = \frac{mdia - mediana}{medida \text{ de escala}}$$

[Morettin and Singer, 2022, p. 67].

A transformação (3.23) [acima] é um caso particular das **transformações de Box-Cox** que são da forma

$$g(x) = \begin{cases} \frac{(x^p - 1)}{p}, & \text{se } p \neq 0 \\ \log(x), & \text{se } p = 0 \end{cases}$$

[Morettin and Singer, 2022, p. 69].

3.11 Notas de capítulo

Sem notas para esta seção.

3.12 Exercícios

Exercício 3.1

O arquivo `rehabcardio.xls` [Carvalho et al., 2007] contém informações sobre um estudo de reabilitação de pacientes cardíacos. Elabore um relatório indicando possíveis inconsistências na matriz de dados e faça uma análise descritiva das variáveis Peso, Altura, Coltot, HDL, LDL, Lesoes, Diabete e HA. Com essa finalidade,

- Construa distribuições de frequências para as variáveis qualitativas.
- Construa histogramas, *boxplots* e gráficos de simetria para as variáveis contínuas.
- Construa uma tabela com medidas resumo para as variáveis contínuas.
- Avalie a compatibilidade de distribuições normais para as variáveis contínuas por meio de gráficos QQ.

Solução.

Preparação dos dados

Vamos iniciar a nossa análise carregando o arquivo com a biblioteca **readxl**.

```
rehabcardio <- readxl::read_xls(paste0(data_dir, "rehabcardio.xls"))
```

Nosso conjunto de dados contém 43 variáveis aferidas em 381 pacientes. Vamos avaliar cada variável individualmente e fazer os ajustes necessários.

As variáveis do conjunto são:

```
colnames(rehabcardio)
```

```
## [1] "Registro"      "Genero"      "Origem"      "Grupo"      "Nascimento"
## [6] "Coleta"        "Programa"    "Lesoes"      "IAM1"        "RM1"
## [11] "ATC1"          "Peso"        "Altura"      "Tabagismo"   "Diabete"
## [16] "HA"            "Coltot"      "HDL"         "LDL"         "Triglic"
## [21] "Glicose"       "Acurico"     "ColtothDL"   "Aspirina"    "Betabloc"
## [26] "Nitrato"       "Hipolip"     "IECA"        "CA"          "Diuretico"
## [31] "Digit"        "Outros"      "Nummed"      "Tempesforco" "Isquemia"
## [36] "RM2"          "DataRM2"     "ATC2"        "DataATC2"    "IAM2"
## [41] "DataIAM2"     "Obito"       "Causa"
```

Um pequeno resumo das variáveis originais é apresentado a seguir:

```
summary(rehabcardio)
```

```
##      Registro      Genero      Origem      Grupo
## Min.   : 1.0      Length:381      Length:381      Length:381
## 1st Qu.:100.0     Class :character  Class :character  Class :character
## Median :197.0     Mode  :character  Mode  :character  Mode  :character
## Mean   :196.3
## 3rd Qu.:293.0
## Max.   :391.0
##
##      Nascimento      Coleta
## Min.   :1913-10-09 00:00:00.000  Min.   :1900-01-01 00:00:00.00
## 1st Qu.:1934-10-01 00:00:00.000  1st Qu.:1900-06-06 00:00:00.00
## Median :1941-07-28 12:00:00.000  Median :1900-08-12 00:00:00.00
## Mean   :1941-10-10 22:32:51.526  Mean   :1902-09-23 16:34:51.31
## 3rd Qu.:1949-05-02 12:00:00.000  3rd Qu.:1900-11-21 00:00:00.00
## Max.   :1976-05-23 00:00:00.000  Max.   :1999-10-27 00:00:00.00
## NA's   :1                      NA's   :9
##      Programa      Lesoes      IAM1
## Min.   :1900-01-17 00:00:00  Min.   :1.000  Min.   :0.0000
## 1st Qu.:1993-07-29 00:00:00  1st Qu.:1.000  1st Qu.:0.0000
## Median :1996-11-21 00:00:00  Median :2.000  Median :1.0000
## Mean   :1980-08-04 05:20:00  Mean   :1.951  Mean   :0.6307
## 3rd Qu.:1998-11-21 18:00:00  3rd Qu.:3.000  3rd Qu.:1.0000
## Max.   :1999-12-10 00:00:00  Max.   :3.000  Max.   :1.0000
## NA's   :3                      NA's   :14  NA's   :10
##      RM1      ATC1      Peso      Altura
## Min.   :0.0000  Min.   :0.0000  Min.   : 47.00  Min.   :1.440
## 1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.: 67.05  1st Qu.:1.630
## Median :0.0000  Median :0.0000  Median : 75.50  Median :1.680
## Mean   :0.4233  Mean   :0.3915  Mean   : 76.11  Mean   :1.675
```

```

## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.: 84.95 3rd Qu.:1.730
## Max. :1.0000 Max. :1.0000 Max. :119.00 Max. :1.910
## NA's :3 NA's :3 NA's :6 NA's :7
## Tabagismo Diabete HA Coltot
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :111.0
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:180.0
## Median :1.0000 Median :0.000 Median :1.0000 Median :206.0
## Mean :0.7204 Mean :0.216 Mean :0.5676 Mean :215.4
## 3rd Qu.:1.0000 3rd Qu.:0.000 3rd Qu.:1.0000 3rd Qu.:238.0
## Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :799.0
## NA's :9 NA's :6 NA's :11 NA's :54
## HDL LDL Triglic Glicose
## Min. : 14.00 Min. : 57.0 Length:381 Min. : 20.0
## 1st Qu.: 35.00 1st Qu.:110.0 Class :character 1st Qu.: 89.0
## Median : 40.00 Median :129.0 Mode :character Median : 97.0
## Mean : 41.52 Mean :137.3 Mean :109.6
## 3rd Qu.: 46.00 3rd Qu.:159.0 3rd Qu.:111.0
## Max. :200.00 Max. :425.0 Max. :379.0
## NA's :65 NA's :92 NA's :81
## Acurico ColtotHDL Aspirina Betabloc
## Min. : 1.800 Min. : 0.000 Min. :0.0000 Min. :0.0000
## 1st Qu.: 4.300 1st Qu.: 4.400 1st Qu.:1.0000 1st Qu.:0.0000
## Median : 5.300 Median : 5.170 Median :1.0000 Median :0.0000
## Mean : 5.567 Mean : 5.425 Mean :0.8223 Mean :0.4868
## 3rd Qu.: 6.600 3rd Qu.: 6.082 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :10.400 Max. :21.071 Max. :1.0000 Max. :1.0000
## NA's :159 NA's :65 NA's :4 NA's :3
## Nitrato Hipolip IECA CA
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
## Median :1.0000 Median :0.0000 Median :0.0000 Median :0.0000
## Mean :0.6138 Mean :0.3583 Mean :0.2888 Mean :0.3687
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000
## NA's :3 NA's :7 NA's :7 NA's :4
## Diuretico Digit Outros Nummed
## Min. :0.00000 Min. :0.000 Min. :0.0000 Min. : 0.000
## 1st Qu.:0.00000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.: 3.000
## Median :0.00000 Median :0.000 Median :1.0000 Median : 4.000
## Mean :0.02133 Mean :0.008 Mean :0.5421 Mean : 3.869
## 3rd Qu.:0.00000 3rd Qu.:0.000 3rd Qu.:1.0000 3rd Qu.: 5.000
## Max. :1.00000 Max. :1.000 Max. :1.0000 Max. :10.000
## NA's :6 NA's :6 NA's :1
## Tempesforco Isquemia RM2
## Min. : 0.00 Min. :0.0000 Min. :0.00000

```

```
## 1st Qu.: 6.00 1st Qu.:0.0000 1st Qu.:0.00000
## Median : 8.00 Median :0.0000 Median :0.00000
## Mean : 7.47 Mean :0.4298 Mean :0.03504
## 3rd Qu.: 9.00 3rd Qu.:1.0000 3rd Qu.:0.00000
## Max. :15.00 Max. :1.0000 Max. :1.00000
## NA's :45 NA's :39 NA's :10
## DataRM2 ATC2
## Min. :1900-08-07 00:00:00 Min. :0.00000
## 1st Qu.:1996-03-05 00:00:00 1st Qu.:0.00000
## Median :1998-06-29 00:00:00 Median :0.00000
## Mean :1989-03-14 00:00:00 Mean :0.08871
## 3rd Qu.:1999-03-17 12:00:00 3rd Qu.:0.00000
## Max. :1999-11-18 00:00:00 Max. :1.00000
## NA's :370 NA's :9
## DataATC2 IAM2
## Min. :1900-01-01 00:00:00.0000 Min. :0.00000
## 1st Qu.:1900-10-16 00:00:00.0000 1st Qu.:0.00000
## Median :1998-04-01 00:00:00.0000 Median :0.00000
## Mean :1970-01-25 01:32:54.1935 Mean :0.06469
## 3rd Qu.:1999-04-01 00:00:00.0000 3rd Qu.:0.00000
## Max. :1999-12-22 00:00:00.0000 Max. :1.00000
## NA's :350 NA's :10
## DataIAM2 Obito
## Min. :1900-02-01 00:00:00.000 Min. :1900-01-26 00:00:00.000
## 1st Qu.:1900-09-30 00:00:00.000 1st Qu.:1900-10-01 00:00:00.000
## Median :1997-03-11 00:00:00.000 Median :1997-03-16 00:00:00.000
## Mean :1965-04-17 18:17:09.571 Mean :1963-09-25 14:07:04.529
## 3rd Qu.:1998-05-01 00:00:00.000 3rd Qu.:1999-02-03 00:00:00.000
## Max. :1999-12-22 00:00:00.000 Max. :1999-12-07 00:00:00.000
## NA's :360 NA's :364
## Causa
## Min. :0.00000
## 1st Qu.:0.00000
## Median :0.00000
## Mean :0.08333
## 3rd Qu.:0.00000
## Max. :2.00000
## NA's :9
```

Vejamos agora os tipos de dados em cada coluna:

```
rehabcardio %>%
  map(typeof) %>%
  as.data.frame() %>%
```



```

t() %>%
kable(
  format = "pipe",
  caption = "**Tabela 3.1:** Tipos de dados das variáveis em `rehabcardio.xls`",
  label = NA,
  digits = 2,
  align = "c",
  format.args = list(decimal.mark = ",")
)

```

Tabela 3.1: Tipos de dados das variáveis em rehabcardio.xls

Registro	double
Genero	character
Origem	character
Grupo	character
Nascimento	double
Coleta	double
Programa	double
Lesoes	double
IAMI	double
RM1	double
ATC1	double
Peso	double
Altura	double
Tabagismo	double
Diabete	double
HA	double
Coltot	double
HDL	double
LDL	double
Triglic	character
Glicose	double
Acurico	double
ColtotHDL	double
Aspirina	double
Betabloc	double
Nitrato	double
Hipolip	double
IECA	double
CA	double
Diuretico	double
Digit	double

Outros	double
Nummed	double
Tempesforco	double
Isquemia	double
RM2	double
DataRM2	double
ATC2	double
DataATC2	double
IAM2	double
DataIAM2	double
Obito	double
Causa	double

Notamos que há vários tipos inconsistentes. Podemos resolver isso informando os tipos durante o carregamento dos dados ou usando a função `mutate()` em conjunção com as funções de `as.*()` do R. Optamos pela segunda forma:

```
rehabcardio <- rehabcardio %>%
  mutate(
    Registro = as.integer(Registro),
    Genero = as.factor(Genero),
    Origem = as.factor(Origem),
    Grupo = as.factor(Grupo),
    Nascimento = as.Date(Nascimento),
    Coleta = as.Date(Coleta),
    Programa = as.Date(Programa),
    Lesoes = as.integer(Lesoes),
    IAM1 = as.logical(IAM1),
    RM1 = as.logical(RM1),
    ATC1 = as.logical(ATC1),
    Peso = as.double(Peso),
    Altura = as.double(Altura),
    Tabagismo = as.logical(Tabagismo),
    Diabetes = as.logical(Diabetes),
    HA = as.logical(HA),
    Coltot = as.double(Coltot),
    HDL = as.double(HDL),
    LDL = as.double(LDL),
    Triglic = as.double(Triglic),
    Glicose = as.double(Glicose),
    Acurico = as.double(Acurico),
    ColtothDL = as.double(ColtothDL),
```

```

    Aspirina = as.logical(Aspirina),
    Betabloc = as.logical(Betabloc),
    Nitrato = as.logical(Nitrato),
    Hipolip = as.logical(Hipolip),
    IECA = as.logical(IECA),
    CA = as.logical(CA),
    Diuretico = as.logical(Diuretico),
    Digit = as.logical(Digit),
    Outros = as.logical(Outros),
    Nummed = as.integer(Nummed),
    Tempesforco = as.double(Tempesforco),
    Isquemia = as.logical(Isquemia),
    RM2 = as.logical(RM2),
    DataRM2 = as.Date(DataRM2),
    ATC2 = as.logical(ATC2),
    DataATC2 = as.Date(DataATC2),
    IAM2 = as.logical(IAM2),
    DataIAM2 = as.Date(DataIAM2),
    Obito = as.Date(Obito),
    Causa = as.integer(Causa)
)

```

```

## Warning: There was 1 warning in `mutate()`.
## i In argument: `Triglic = as.double(Triglic)`.
## Caused by warning:
## ! NAs introduzidos por coerção

```

Antes de seguirmos, vamos salvar nosso conjunto de dados em um novo arquivo para facilitar análises futuras.

```

rehabcardio %>%
  write_csv(paste0(data_dir, "rehabcardio.csv"))

```

Podemos agora seguir com a análise dos dados.

Distribuição de frequência para as variáveis qualitativas

As tabelas 3.2 e 3.3 a seguir apresentam a distribuição de frequência para as variáveis Diabete e HA.

```

rehabcardio %>%
  group_by(Diabete) %>%
  summarize(
    `Frequência observada` = n(),
    `Frequência relativa` = 100 * n() / nrow(rehabcardio)
  ) %>%
  mutate(
    `Frequência acumulada` = cumsum(`Frequência relativa`),
    Diabete = case_when(
      Diabete ~ "Presente",
      !Diabete ~ "Ausente",
      is.na(Diabete) ~ "Sem resposta"
    )
  ) %>%
  bind_rows(
    tribble(
      ~Diabete, ~`Frequência observada`, ~`Frequência relativa`, ~`Frequência acumulada`,
      "**Total**", nrow(rehabcardio), 100, 100
    )
  ) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.2:** Distribuição de frequência da variável `Diabete`",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )

```

Tabela 3.2: Distribuição de frequência da variável `Diabete`

Diabete	Frequência observada	Frequência relativa	Frequência acumulada
Ausente	294	77,17	77,17
Presente	81	21,26	98,43
Sem resposta	6	1,57	100,00
Total	381	100,00	100,00

```

rehabcardio %>%
  group_by(HA) %>%

```

```

summarize(
  `Frequência observada` = n(),
  `Frequência relativa` = 100 * n() / nrow(rehabcardio)
) %>%
mutate(
  `Frequência acumulada` = cumsum(`Frequência relativa`),
  `Diagnóstico` = case_when(
    HA ~ "Hipertenso",
    !HA ~ "Normotenso",
    is.na(HA) ~ "Sem resposta"
  )
) %>%
bind_rows(
  tribble(
    ~`Diagnóstico`, ~`Frequência observada`, ~`Frequência relativa`, ~`Frequência acumulada`,
    "**Total**", nrow(rehabcardio), 100, 100
  )
) %>%
select(`Diagnóstico`, `Frequência observada`, `Frequência relativa`, `Frequência acumulada`) %>%
kable(
  format = "pipe",
  caption = "**Tabela 3.3:** Distribuição de frequência da variável `HA`",
  label = NA,
  digits = 2,
  align = "c",
  format.args = list(decimal.mark = ",")
)

```

Tabela 3.3: Distribuição de frequência da variável HA

Diagnóstico	Frequência observada	Frequência relativa	Frequência acumulada
Normotenso	160	41,99	41,99
Hipertenso	210	55,12	97,11
Sem resposta	11	2,89	100,00
Total	381	100,00	100,00

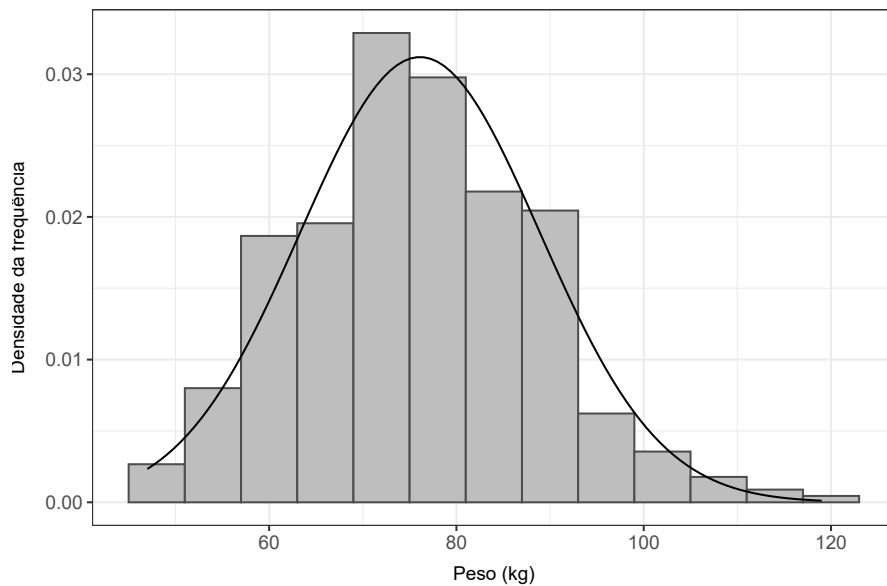
Histogramas, boxplots e gráficos de simetria para as variáveis contínuas

```
rehabcardio %>%  
  ggplot(aes(x = Peso)) +  
    geom_histogram(  
      mapping = aes(y = ..density..),  
      binwidth = 6,  
      fill = "grey",  
      color = "grey30"  
    ) +  
    geom_function(  
      fun = dnorm,  
      args = list(  
        mean = mean(rehabcardio$Peso, na.rm = TRUE),  
        sd = sd(rehabcardio$Peso, na.rm = TRUE)  
      )  
    ) +  
    labs(  
      title = "Figura 3.1: Distribuição de frequência dos pesos dos pacientes",  
      x = "Peso (kg)",  
      y = "Densidade da frequência"  
    ) +  
    tema
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## i Please use `after_stat(density)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was  
## generated.
```

```
## Warning: Removed 6 rows containing non-finite values (`stat_bin()`).
```

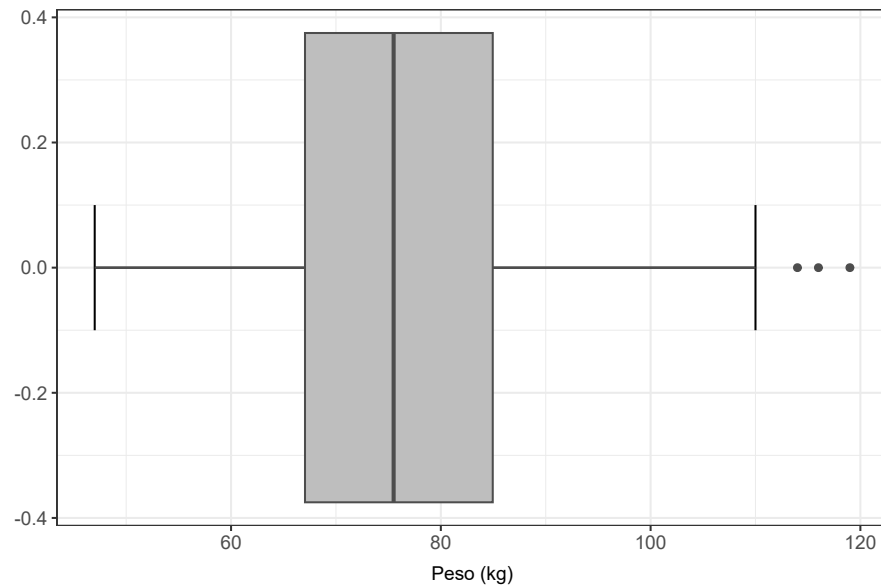
Figura 3.1: Distribuição de frequência dos pesos dos pacientes



```
rehabcardio %>%
  ggplot(aes(x = Peso)) +
    stat_boxplot(geom = "errorbar", width = 0.2) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.2: Distribuição do peso dos pacientes",
      x = "Peso (kg)"
    ) +
    tema
```

```
## Warning: Removed 6 rows containing non-finite values (`stat_boxplot()`).
## Removed 6 rows containing non-finite values (`stat_boxplot()`).
```

Figura 3.2: Distribuição do peso dos pacientes



```
geom_symmetry <- function(x) {
  x <- sort(x)

  mediana <- median(x)
  n <- length(x)
  meio_n <- (n+1)/2

  uv = tibble(
    u = rep(0, meio_n),
    v = rep(0, meio_n)
  )

  for (i in seq(1, meio_n, 1)) {
    uv$u[i] <- mediana - x[i]
    uv$v[i] <- x[n + 1 - i] - mediana
  }

  geom_point(data = uv, mapping = aes(u, v))
}

rehabcardio %>%
  ggplot() +
  geom_symmetry(rehabcardio$Peso) +
```



```
geom_abline(aes(intercept = 0, slope = 1), color = "red", alpha = .5) +
labs(
  title = "Figura 3.3: Gráfico de simetria para o peso dos pacientes",
  x = "Dispersão inferior",
  y = "Dispersão superior"
) +
tema
```

Figura 3.3: Gráfico de simetria para o peso dos pacientes



```
rehabcardio %>%
  ggplot(aes(x = Altura)) +
  geom_histogram(
    mapping = aes(y = ..density..),
    binwidth = 0.06,
    fill = "grey",
    color = "grey30"
  ) +
  geom_function(
    fun = dnorm,
    args = list(
      mean = mean(rehabcardio$Altura, na.rm = TRUE),
      sd = sd(rehabcardio$Altura, na.rm = TRUE)
    )
  )
```

```

) +
labs(
  title = "Figura 3.4: Distribuição de frequência das alturas dos pacientes",
  x = "Altura (m)",
  y = "Densidade da frequência"
) +
tema

```

```
## Warning: Removed 7 rows containing non-finite values (`stat_bin()`).
```



```

rehabcardio %>%
  ggplot(aes(x = Altura)) +
    stat_boxplot(geom = "errorbar", width = 0.2) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.5: Distribuição da altura dos pacientes",
      x = "Altura (m)"
    ) +
    tema

```

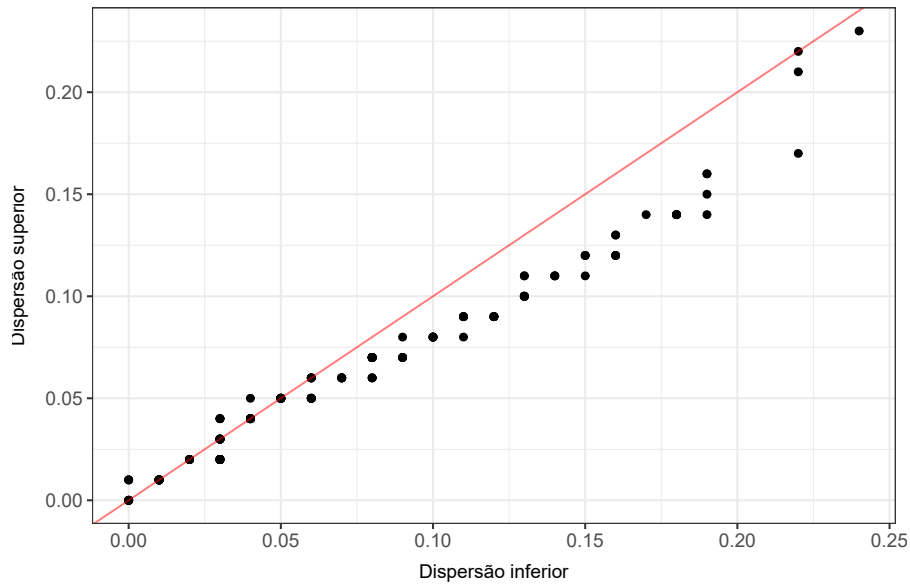
```
## Warning: Removed 7 rows containing non-finite values (`stat_boxplot()`).  
## Removed 7 rows containing non-finite values (`stat_boxplot()`).
```

Figura 3.5: Distribuição da altura dos pacientes



```
rehabcardio %>%  
  ggplot() +  
    geom_symmetry(rehabcardio$Altura) +  
    geom_abline(aes(intercept = 0, slope = 1), color = "red", alpha = .5) +  
    labs(  
      title = "Figura 3.6: Gráfico de simetria para a altura dos pacientes",  
      x = "Dispersão inferior",  
      y = "Dispersão superior"  
    ) +  
    tema
```

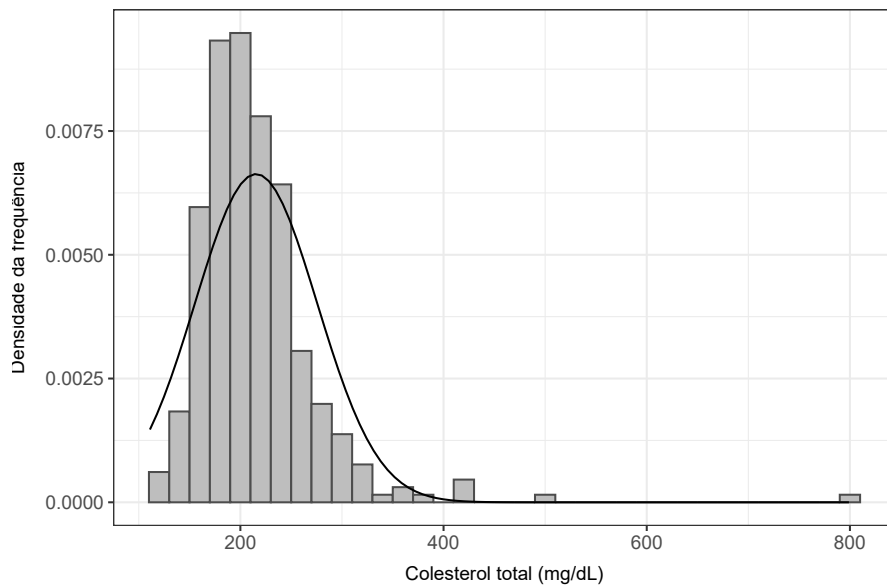
Figura 3.6: Gráfico de simetria para a altura dos pacientes



```
rehabcardio %>%
  ggplot(aes(x = Coltot)) +
    geom_histogram(
      mapping = aes(y = ..density..),
      binwidth = 20,
      fill = "grey",
      color = "grey30"
    ) +
    geom_function(
      fun = dnorm,
      args = list(
        mean = mean(rehabcardio$Coltot, na.rm = TRUE),
        sd = sd(rehabcardio$Coltot, na.rm = TRUE)
      )
    ) +
    labs(
      title = "Figura 3.7: Distribuição de frequência do colesterol total dos pacientes",
      x = "Colesterol total (mg/dL)",
      y = "Densidade da frequência"
    ) +
    tema
```

```
## Warning: Removed 54 rows containing non-finite values (`stat_bin()`).
```

Figura 3.7: Distribuição de frequência do colesterol total dos pacientes



```
rehabcardio %>%  
  ggplot(aes(x = Coltot)) +  
    stat_boxplot(geom = "errorbar", width = 0.2) +  
    geom_boxplot(  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.9: Distribuição do colesterol total dos pacientes",  
      x = "Colesterol total (mg/dL)"  
    ) +  
    tema
```

```
## Warning: Removed 54 rows containing non-finite values (`stat_boxplot()`).  
## Removed 54 rows containing non-finite values (`stat_boxplot()`).
```

Figura 3.9: Distribuição do colesterol total dos pacientes



```
rehabcardio %>%
  ggplot() +
    geom_symmetry(rehabcardio$Coltot) +
    geom_abline(aes(intercept = 0, slope = 1), color = "red", alpha = .5) + labs(
      title = "Figura 3.10: Gráfico de simetria para os níveis de colesterol total dos pacientes",
      x = "Dispersão inferior",
      y = "Dispersão superior"
    ) +
    tema
```

Figura 3.10: Gráfico de simetria para os níveis de colesterol total dos pacientes



```
rehabcardio %>%
  ggplot(aes(x = HDL)) +
    geom_histogram(
      mapping = aes(y = ..density..),
      binwidth = 7,
      fill = "grey",
      color = "grey30"
    ) +
    geom_function(
      fun = dnorm,
      args = list(
        mean = mean(rehabcardio$HDL, na.rm = TRUE),
        sd = sd(rehabcardio$HDL, na.rm = TRUE)
      )
    ) +
    labs(
      title = "Figura 3.11: Distribuição de frequência do colesterol HDL dos pacientes",
      x = "Colesterol HDL (mg/dL)",
      y = "Densidade da frequência"
    ) +
    tema
```

```
## Warning: Removed 65 rows containing non-finite values (`stat_bin()`).
```

Figura 3.11: Distribuição de frequência do colesterol HDL dos pacientes



```
rehabcardio %>%  
  ggplot(aes(x = HDL)) +  
    stat_boxplot(geom = "errorbar", width = 0.2) +  
    geom_boxplot(  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.12: Distribuição do colesterol HDL dos pacientes",  
      x = "Colesterol HDL (mg/dL)"  
    ) +  
    tema
```

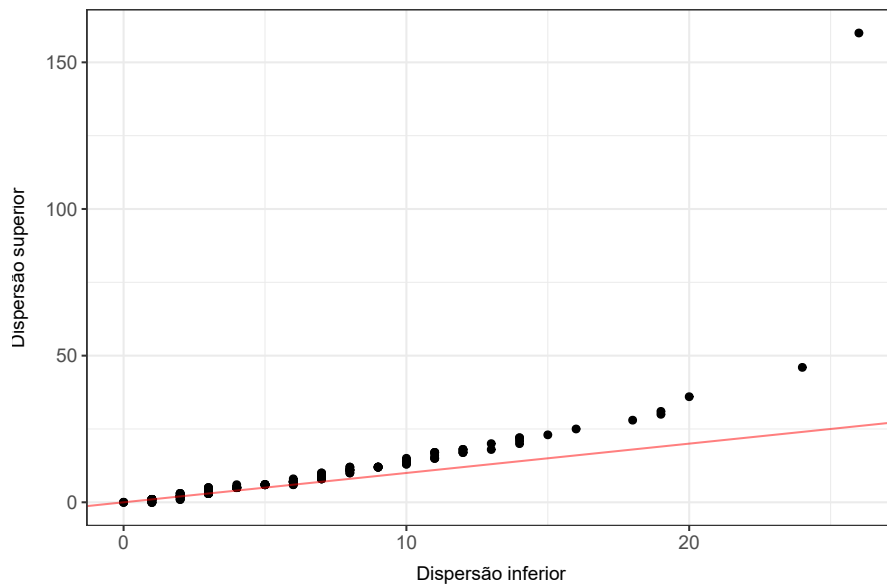
```
## Warning: Removed 65 rows containing non-finite values (`stat_boxplot()`).  
## Removed 65 rows containing non-finite values (`stat_boxplot()`).
```


Figura 3.12: Distribuição do colesterol HDL dos pacientes



```
rehabcardio %>%
  ggplot() +
    geom_symmetry(rehabcardio$HDL) +
    geom_abline(aes(intercept = 0, slope = 1), color = "red", alpha = .5) +
    labs(
      title = "Figura 3.13: Gráfico de simetria para os níveis de colesterol HDL dos pacientes",
      x = "Dispersão inferior",
      y = "Dispersão superior"
    ) +
    tema
```

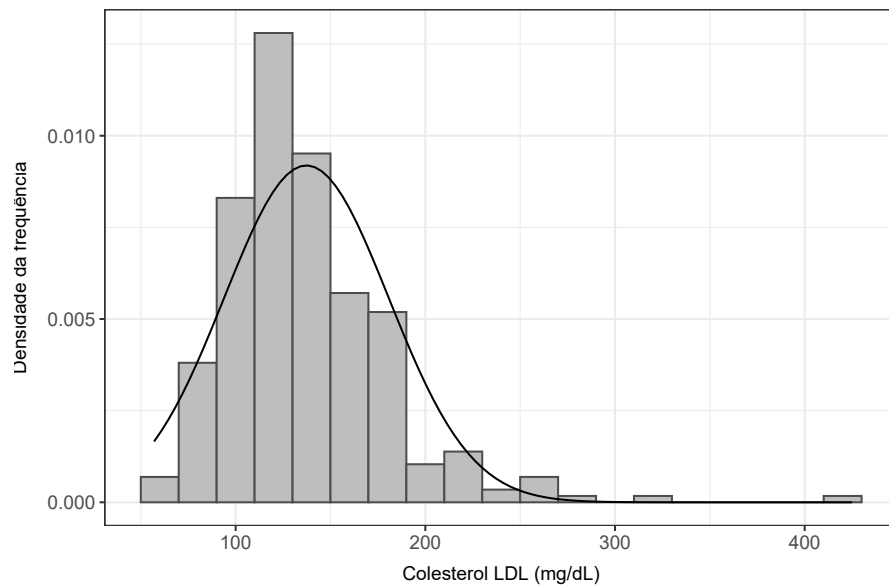
Figura 3.13: Gráfico de simetria para os níveis de colesterol HDL dos pacientes



```
rehabcardio %>%
  ggplot(aes(x = LDL)) +
    geom_histogram(
      mapping = aes(y = ..density..),
      binwidth = 20,
      fill = "grey",
      color = "grey30"
    ) +
    geom_function(
      fun = dnorm,
      args = list(
        mean = mean(rehabcardio$LDL, na.rm = TRUE),
        sd = sd(rehabcardio$LDL, na.rm = TRUE)
      )
    ) +
    labs(
      title = "Figura 3.14: Distribuição de frequência do colesterol LDL dos pacientes",
      x = "Colesterol LDL (mg/dL)",
      y = "Densidade da frequência"
    ) +
    tema
```

```
## Warning: Removed 92 rows containing non-finite values (`stat_bin()`).
```

Figura 3.14: Distribuição de frequência do colesterol LDL dos pacientes



```
rehabcardio %>%  
  ggplot(aes(x = LDL)) +  
    stat_boxplot(geom = "errorbar", width = 0.2) +  
    geom_boxplot(  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.15: Distribuição do colesterol LDL dos pacientes",  
      x = "Colesterol LDL (mg/dL)"  
    ) +  
    tema
```

```
## Warning: Removed 92 rows containing non-finite values (`stat_boxplot()`).  
## Removed 92 rows containing non-finite values (`stat_boxplot()`).
```

Figura 3.15: Distribuição do colesterol LDL dos pacientes



```
rehabcardio %>%
  ggplot() +
    geom_symmetry(rehabcardio$LDL) +
    geom_abline(aes(intercept = 0, slope = 1), color = "red", alpha = .5) +
    labs(
      title = "Figura 3.16: Gráfico de simetria para os níveis de colesterol LDL dos pacientes",
      x = "Dispersão inferior",
      y = "Dispersão superior"
    ) +
    tema
```

Figura 3.16: Gráfico de simetria para os níveis de colesterol LDL dos pacientes



c) Construa uma tabela com medidas resumo para as variáveis contínuas

```
summ <- function(x) {
  c(
    n = sum(!is.na(x), na.rm = TRUE),
    min = min(x, na.rm = TRUE),
    max = max(x, na.rm = TRUE),
    mean = mean(x, na.rm = TRUE),
    sd = sd(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
    iqr = IQR(x, na.rm = TRUE)
  )
}

resumo <- rehabcardio %>%
  select(Peso, Altura, Coltot, HDL, LDL) %>%
  map(summ) %>%
  as.data.frame() %>%
  t()

colnames(resumo) <- c("Quantidade de observações", "Mínimo", "Máximo", "Média", "Desvio padrão", "Mediana", "Dispersão inferior", "Dispersão superior")
```

```

resumo %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.4:** Medidas de resumo para algumas variáveis de `rehabcardio.xls`",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )

```

Tabela 3.4: Medidas de resumo para algumas variáveis de rehabcardio.xls

	Quantidade de observações	Mínimo	Máximo	Média	Desvio padrão	Mediana	Distância interquartil
Peso	375	47,00	119,00	76,11	12,79	75,50	17,9
Altura	374	1,44	1,91	1,67	0,08	1,68	0,1
Coltot	327	111,00	799,00	215,43	60,16	206,00	58,0
HDL	316	14,00	200,00	41,52	13,14	40,00	11,0
LDL	289	57,00	425,00	137,31	43,41	129,00	49,0

d) Avalie a compatibilidade de distribuições normais para as variáveis contínuas por meio de gráficos QQ

```

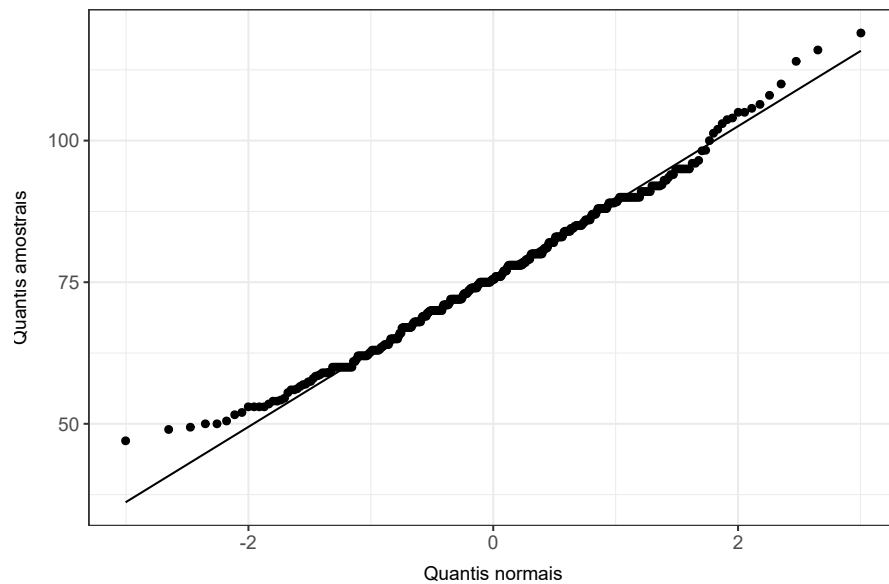
rehabcardio %>%
  ggplot(aes(sample = Peso)) +
    geom_qq() +
    geom_qq_line() +
    labs(
      title = "Figura 3.17: Gráfico QQ normal para o peso dos pacientes",
      x = "Quantis normais",
      y = "Quantis amostrais"
    ) +
    tema

## Warning: Removed 6 rows containing non-finite values (`stat_qq()`).

## Warning: Removed 6 rows containing non-finite values (`stat_qq_line()`).

```

Figura 3.17: Gráfico QQ normal para o peso dos pacientes



```
rehabcardio %>%  
  ggplot(aes(sample = Altura)) +  
    geom_qq() +  
    geom_qq_line() +  
    labs(  
      title = "Figura 3.18: Gráfico QQ normal para a altura dos pacientes",  
      x = "Quantis normais",  
      y = "Quantis amostrais"  
    ) +  
    tema
```

```
## Warning: Removed 7 rows containing non-finite values (`stat_qq()`).
```

```
## Warning: Removed 7 rows containing non-finite values (`stat_qq_line()`).
```

Figura 3.18: Gráfico QQ normal para a altura dos pacientes



```
rehabcardio %>%  
  ggplot(aes(sample = Coltot)) +  
    geom_qq() +  
    geom_qq_line() +  
    labs(  
      title = "Figura 3.19: Gráfico QQ normal para os níveis de colesterol total dos pacientes",  
      x = "Quantis normais",  
      y = "Quantis amostrais"  
    ) +  
    tema
```

```
## Warning: Removed 54 rows containing non-finite values (`stat_qq()`).
```

```
## Warning: Removed 54 rows containing non-finite values (`stat_qq_line()`).
```


Figura 3.19: Gráfico QQ normal para os níveis de colesterol total dos pacientes



```
rehabcardio %>%  
  ggplot(aes(sample = HDL)) +  
    geom_qq() +  
    geom_qq_line() +  
    labs(  
      title = "Figura 3.20: Gráfico QQ normal para os níveis de colesterol HDL dos pacientes",  
      x = "Quantis normais",  
      y = "Quantis amostrais"  
    ) +  
    tema
```

```
## Warning: Removed 65 rows containing non-finite values (`stat_qq()`).
```

```
## Warning: Removed 65 rows containing non-finite values (`stat_qq_line()`).
```

Figura 3.20: Gráfico QQ normal para os níveis de colesterol HDL dos pacientes



```
rehabcardio %>%  
  ggplot(aes(sample = LDL)) +  
    geom_qq() +  
    geom_qq_line() +  
    labs(  
      title = "Figura 3.21: Gráfico QQ normal para os níveis de colesterol LDL dos pacientes",  
      x = "Quantis normais",  
      y = "Quantis amostrais"  
    ) +  
    tema
```

```
## Warning: Removed 92 rows containing non-finite values (`stat_qq()`).
```

```
## Warning: Removed 92 rows containing non-finite values (`stat_qq_line()`).
```

Figura 3.21: Gráfico QQ normal para os níveis de colesterol LDL dos pacientes



Exercício 3.2

Considere os dados do arquivo `antracose.xls`.

- Construa uma tabela com as medidas de posição e dispersão estudadas para as variáveis desse arquivo.
- Construa histogramas e boxplots para essas variáveis e verifique que transformação é necessária para tornar mais simétricas aquelas em que a simetria pode ser questionada.

Solução. Vamos começar carregando o arquivo.

```
antracose <- readxl::read_xls(paste0(data_dir, "antracose.xls"))
```

Construa uma tabela com as medidas de posição e dispersão estudadas para as variáveis desse arquivo

A Tabela 3.5 apresenta as medidas de posição e dispersão dos dados.

```

resumo <- antracose %>%
  map(summ) %>%
  as.data.frame() %>%
  t()

colnames(resumo) <- c("Quantidade de observações", "Mínimo", "Máximo", "Média", "Desvio padrão", "Mediana", "Distância interquartil")

resumo %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.5:** Medidas de resumo para algumas variáveis de `antracose.xls`",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )

```

Tabela 3.5: Medidas de resumo para algumas variáveis de antracose.xls

	Quantidade de observações	Mínimo	Máximo	Média	Desvio padrão	Mediana	Distância interquartil
idade	2452	19,00	110,00	68,43	15,14	69,00	23,00
tmunic	2452	0,00	103,00	49,27	20,63	50,00	23,00
htransp	2452	0,00	12,00	1,78	2,54	0,83	2,00
cargatabag	2452	0,00	256,00	25,57	39,31	1,50	41,00
antracose	2350	0,00	0,83	0,20	0,16	0,17	0,22
ses	2452	-	1,00	-	0,36	-0,31	0,47
		1,00		0,24			
densid	2452	0,00	0,04	0,02	0,01	0,02	0,01
distmin	2452	0,03	1964,02	170,67	190,87	114,12	180,71

Construa histogramas e boxplots para essas variáveis e verifique que transformação é necessária para tornar mais simétricas aquelas em que a simetria pode ser questionada

Para a variável `idade`, temos o seguinte:

```

antracose %>%
  ggplot(aes(x = idade)) +
  geom_histogram(
    binwidth = 3,

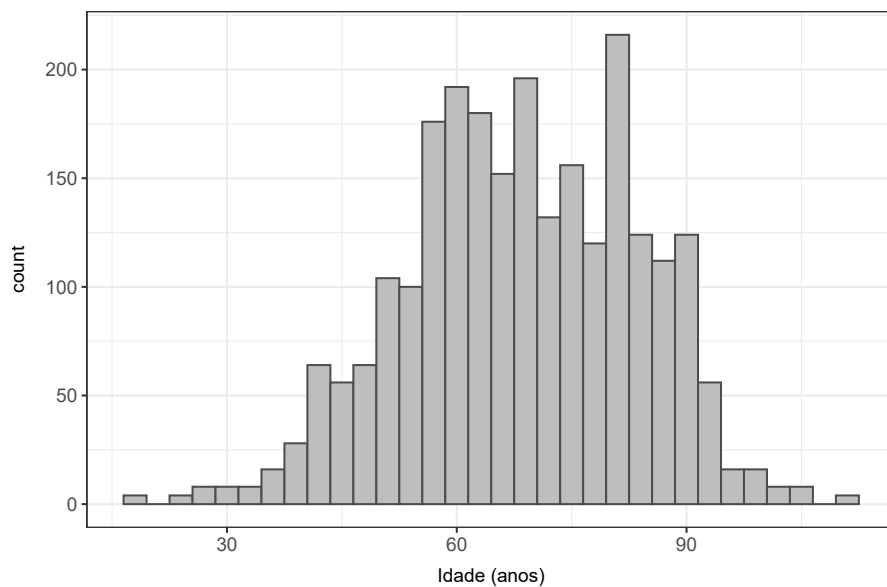
```

```

    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = "Figura 3.22: Distribuição da idade dos pacientes",
    x = "Idade (anos)"
  ) +
  tema

```

Figura 3.22: Distribuição da idade dos pacientes

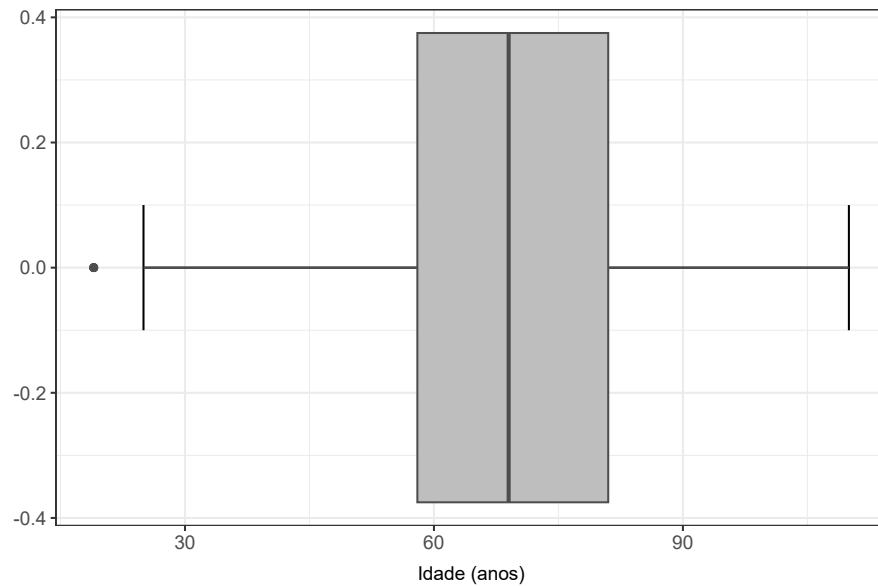


```

antracose %>%
  ggplot(aes(x = idade)) +
  stat_boxplot(geom = "errorbar", width = 0.2) +
  geom_boxplot(
    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = "Figura 3.23: Idade dos pacientes",
    x = "Idade (anos)"
  ) +
  tema

```

Figura 3.23: Idade dos pacientes



Observando os gráficos, podemos perceber uma leve (muito leve) assimetria à esquerda. O coeficiente de simetria de Fisher-Pearson ajustado é negativo (-0.1797772), o que indica a leve assimetria.

Entendemos que não seria necessário nenhuma transformação, uma vez que a assimetria é muito leve, porém se usarmos a transformação de Box-Cox, teremos o seguinte:

```
p <- seq(-3,3, 1/10) # c(-3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3)

transformar <- function(x, p) {
  if(p > 0) {
    x^p
  } else if (p == 0) {
    log(x)
  } else {
    -x^p
  }
}

boxcox <- function(x, p) {
  if(p == 0) {
    log(x)
  } else {
    (((x ^ p) - 1) / p)
  }
}
```

```

    }
  }

  calcular_metrica <- function(x) {
    x <- x[is.finite(x)]
    (mean(x, na.rm = TRUE) - median(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
  }

  dps <- vector("double", length(p))
  for (i in seq_along(p)) {
    dps[[i]] <- antracose$idade %>%
      transformar(p[i]) %>%
      calcular_metrica()
  }

  p[min_rank(dps)[1]]

## [1] -1

```

Após a transformação, verificação que o valor que minimiza a nossa métrica de transformação é $p = -1$, que faz a transformação levar x no seu simétrico $-x$. Isso corrobora com a nossa hipótese de que a função já é a mais simétrica possível (nos termos da nossa transformação).

Para a variável `tmunic`, temos:

```

antracose %>%
  ggplot(aes(x = tmunic)) +
    geom_histogram(
      binwidth = 8,
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.24: Distribuição da variável `tmunic`",
      y = "Frequência"
    ) +
    tema

```

Figura 3.24: Distribuição da variável `tmunic`



```
antracose %>%  
  ggplot(aes(x = tmunic)) +  
    stat_boxplot(geom = "errorbar", width = 0.2) +  
    geom_boxplot(  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.25: Distribuição da variável `tmunic`"  
    ) +  
    tema
```


Figura 3.25: Distribuição da variável `tmunic`



Observando os gráficos, podemos perceber uma leve (muito leve) assimetria à esquerda. O coeficiente de simetria de Fisher-Pearson ajustado é negativo (-0.3960939), o que indica a leve assimetria.

Podemos usar uma transformação para corrigir a assimetria dos dados:

```
dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[[i]] <- antracose$tmunic %>%
    boxcox(p[i]) %>%
    calcular_metrica()
}

p[min_rank(dps)[1]]
```

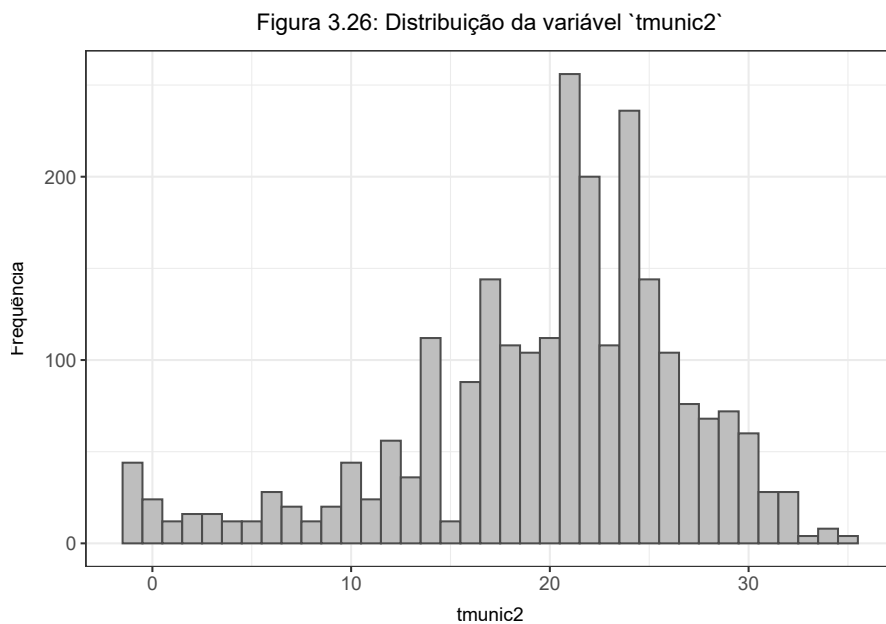
```
## [1] 0.7
```

Notamos que o melhor valor de p é $-1/3$. Com isso, podemos realizar utilizar esta transformação.

```
antracose$tmunic2 <- boxcox(antracose$tmunic, p[min_rank(dps)[1]])

antracose %>%
```

```
ggplot(aes(x = tmunic2)) +
  geom_histogram(
    binwidth = 1,
    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = "Figura 3.26: Distribuição da variável `tmunic2`",
    y = "Frequência"
  ) +
  tema
```

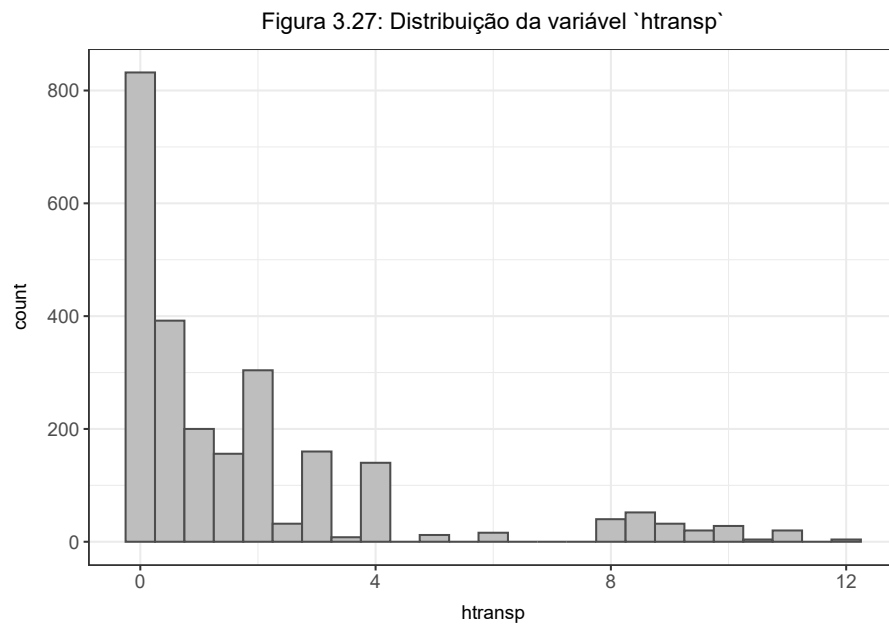


Neste caso, a transformação não melhorou a simetria dos dados. Em vez disso, o coeficiente de assimetria saltou de -0.3960939 para -0.9443456.

Para a variável `htransp`, temos:

```
antracose %>%
  ggplot(aes(x = htransp)) +
  geom_histogram(
    binwidth = 0.5,
    fill = "grey",
    color = "grey30"
  ) +
```

```
labs(
  title = "Figura 3.27: Distribuição da variável `htransp`"
) +
tema
```



```
antracose %>%
  ggplot(aes(x = htransp)) +
  stat_boxplot(geom = "errorbar", width = 0.2) +
  geom_boxplot(
    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = "Figura 3.28: Distribuição da variável `htransp`"
  ) +
tema
```

Figura 3.28: Distribuição da variável 'htransp'



A variável contém uma forte assimetria a direita, conforme atesta também o coeficiente de simetria de Fisher-Pearson ajustado (2.0507871).

Vamos realizar uma transformação:

```
dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[[i]] <- antracose$htransp %>%
    boxcox(p[i]) %>%
    calcular_metrica()
}

p[min_rank(dps)[i]]
```

```
## [1] 0.9
```

Para os valores de p , o melhor ajuste se dá quando $p = 0.9$, vamos utilizá-lo então!

```
antracose$htransp2 <- boxcox(antracose$htransp, p[min_rank(dps)[i]])

antracose %>%
  ggplot(aes(x = htransp2)) +
  geom_histogram(
```

```
binwidth = .5,  
fill = "grey",  
color = "grey30"  
) +  
labs(  
  title = "Figura 3.29: Distribuição da variável `log(htransp)`"  
) +  
tema
```



Isso melhora um pouco a simetria dos dados, mas ainda assim não é suficiente. Vamos por exemplo, comparar os valores antes e depois da transformação.

```
e1071::skewness(antracose$htransp, type = 2)
```

```
## [1] 2.050787
```

```
e1071::skewness(antracose$htransp2, type = 2)
```

```
## [1] 1.851695
```

Vejamos agora a variável `cargatabag`:

```
antracose %>%  
  ggplot(aes(x = cargatabag)) +  
    geom_histogram(  
      binwidth = 15,  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.30: Distribuição da variável `cargatabag`"  
    ) +  
    tema
```

Figura 3.30: Distribuição da variável ``cargatabag``



```
antracose %>%  
  ggplot(aes(x = cargatabag)) +  
    stat_boxplot(geom = "errorbar", width = 0.2) +  
    geom_boxplot(  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(
```

```
title = "Figura 3.31: Distribuição da variável `cargatabag`"
) +
tema
```



Temos uma distribuição fortemente assimétrica à direita. O Coeficiente de Assimetria de Fisher-Pearson Ajustado nos indica um grau de assimetria na ordem de 2.0334838.

Como a distribuição tem um formato exponencial, apostaremos que a transformação que leva x em $\log(x)$ seja suficiente para melhorar a distribuição dos dados.

```
antracose %>%
  ggplot(aes(log(cargatabag))) +
  geom_histogram(
    binwidth = .6,
    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = "Figura 32: Distribuição da variável `cargatabag`"
  ) +
tema
```

```
## Warning: Removed 1196 rows containing non-finite values (`stat_bin()`).
```

Figura 32: Distribuição da variável `cargatabag`



Continuamos com assimetria, desta vez à esquerda. Vamos realizar outras transformações e avaliar a nossa métrica.

```
dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[[i]] <- antracose$htransp %>%
    boxcox(p[i]) %>%
    calcular_metrica()
}

p[min_rank(dps)[1]]
```

```
## [1] 0.3
```

Notamos que para $p = 0.3$, nossa métrica é minimizada.

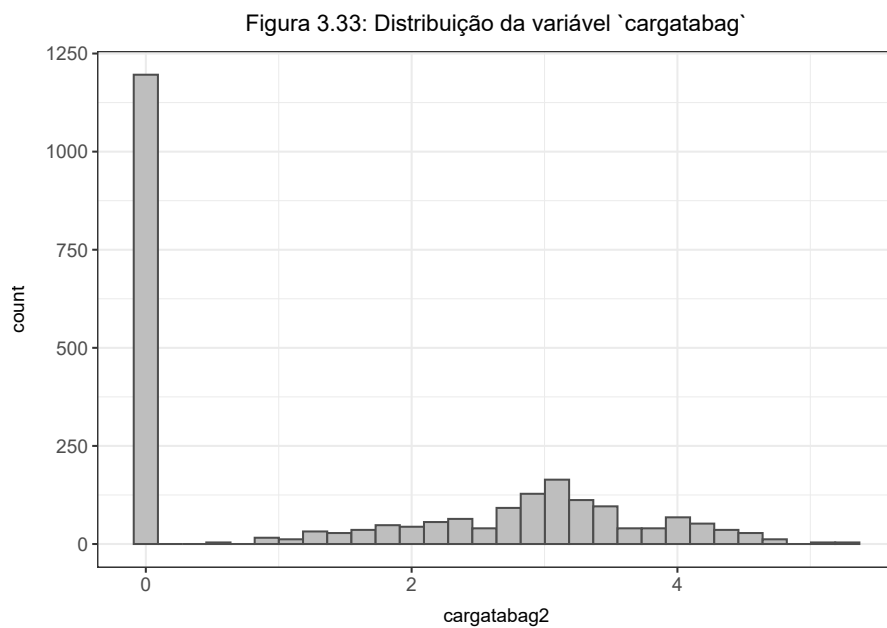
```
antracose$cargatabag2 <- transformar(antracose$cargatabag, p[min_rank(dps)[1]])

antracose %>%
  ggplot(aes(x = cargatabag2)) +
    geom_histogram(
      fill = "grey",
```



```
color = "grey30"  
) +  
labs(  
  title = "Figura 3.33: Distribuição da variável `cargatabag`"  
) +  
tema
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Note que, os valores continuam concentrados no zero, mas os demais, se distribuem de forma mais simétrica do que antes. É importante destacar que, nem sempre será possível tornar simétrica uma distribuição. Neste caso, deveremos procurar técnicas que não dependam da simetria dos dados.

Analisemos agora a variável `antracose`:

```
antracose %>%  
  ggplot(aes(x = antracose)) +  
  geom_histogram(  
    binwidth = .05,  
    fill = "grey",  
    color = "grey30"
```

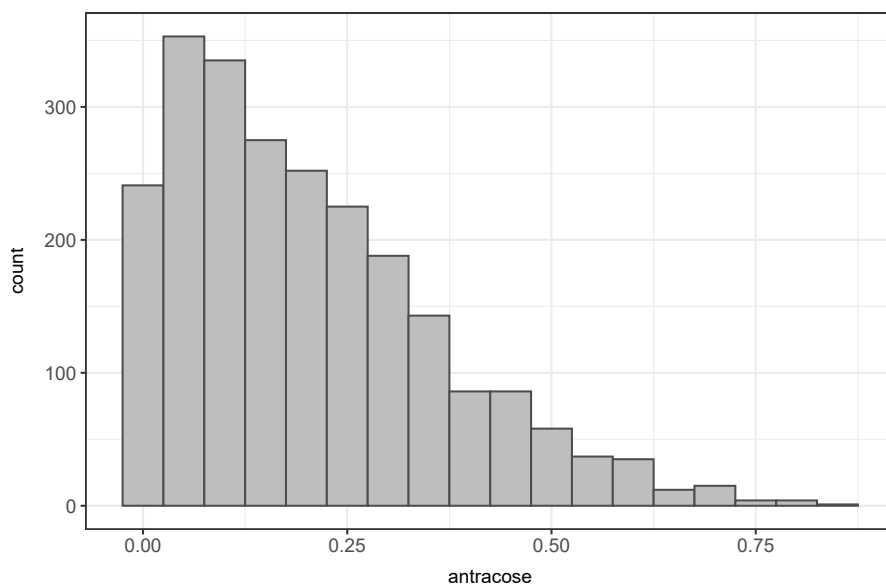
```

) +
labs(
  title = "Figura 3.34: Distribuição da variável `antracose`"
) +
tema

```

```
## Warning: Removed 102 rows containing non-finite values (`stat_bin()`).
```

Figura 3.34: Distribuição da variável `antracose`



```

antracose %>%
  ggplot(aes(x = antracose)) +
    stat_boxplot(geom = "errorbar", width = 0.2) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.35: Distribuição da variável `antracose`"
    ) +
    tema

```

```
## Warning: Removed 102 rows containing non-finite values (`stat_boxplot()`).
```

```
## Removed 102 rows containing non-finite values (`stat_boxplot()`).
```

Figura 3.35: Distribuição da variável `antracose`



Temos uma forte assimetria à direita. O Coeficiente de Assimetria de Fisher-Pearson Ajustado é NA. Vamos verificar se podemos usar uma transformação nos dados.

```
dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[[i]] <- antracose$antracose %>%
    boxcox(p[i]) %>%
    calcular_metrica()
}

p[min_rank(dps)[1]]
```

```
## [1] 0.3
```

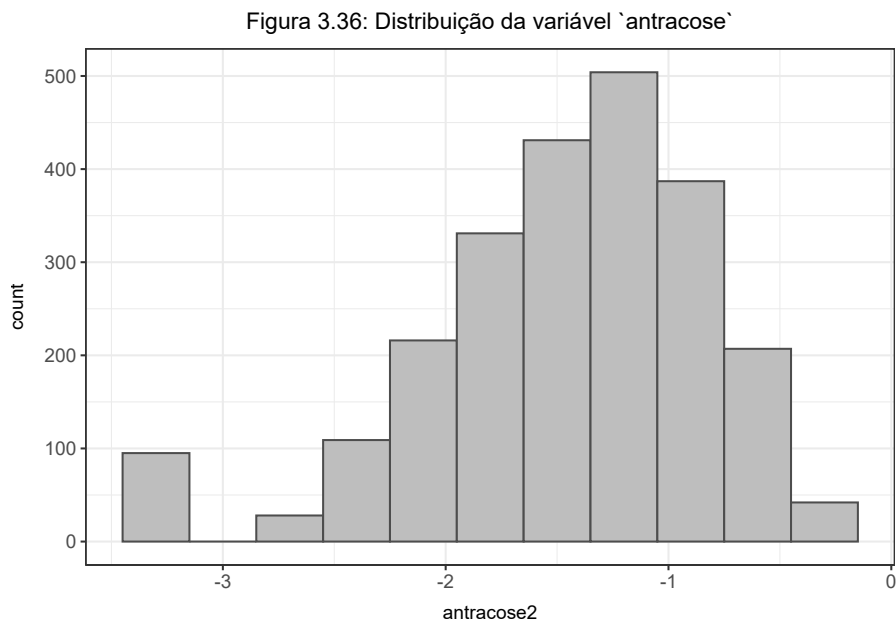
O valor $p = 0.3$ minimiza a nossa medida de avaliação. Vamos agora plotar os dados transformados e avaliar novamente o coeficiente de assimetria.

```
antracose$antracose2 <- boxcox(antracose$antracose, p[min_rank(dps)[1]])

antracose %>%
  ggplot(aes(x = antracose2)) +
  geom_histogram(
```

```
binwidth = .3,  
fill = "grey",  
color = "grey30"  
) +  
labs(  
  title = "Figura 3.36: Distribuição da variável `antracose`"  
) +  
tema
```

```
## Warning: Removed 102 rows containing non-finite values (`stat_bin()`).
```



Parece-nos que os dados se tornaram um pouco mais simétricos, mas nada muito significativo. Talvez valha a pena entender se a concentração dos valores em zero é uma medida real ou se apenas serve para simbolizar, por exemplo, um valor que não pode ser medido devido ao valor de outra variável.

```
e1071::skewness(antracose$antracose, na.rm = TRUE, type = 2)
```

```
## [1] 0.9535379
```

```
e1071::skewness(antracose$antracose2, na.rm = TRUE, type = 2)
```

```
## [1] -0.9171354
```

Para a variável `ses`, temos:

```
antracose %>%  
  ggplot(aes(x = ses)) +  
    geom_histogram(  
      binwidth = .2,  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.37: Distribuição da variável `ses`"  
    ) +  
    tema
```

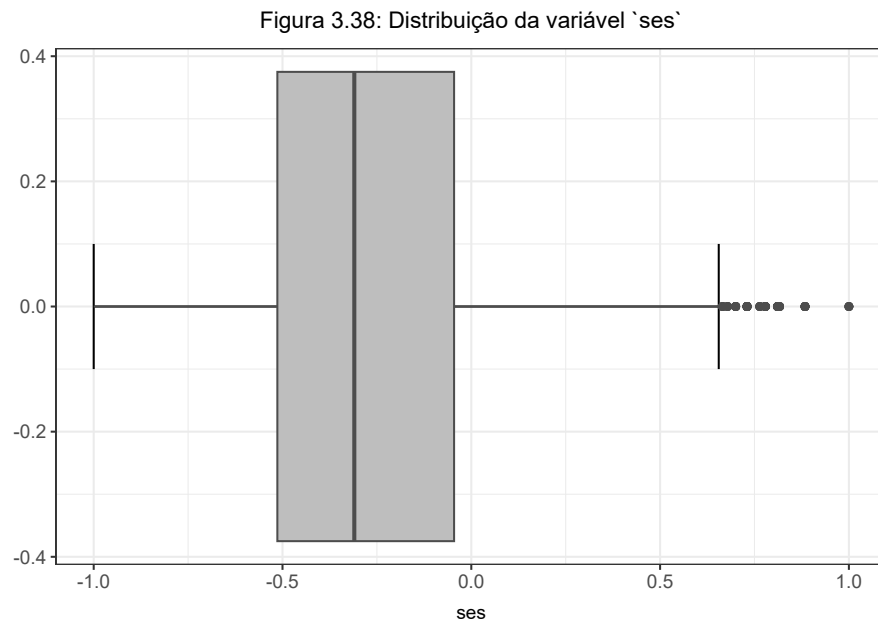


```
antracose %>%  
  ggplot(aes(x = ses)) +
```

```

stat_boxplot(geom = "errorbar", width = 0.2) +
geom_boxplot(
  fill = "grey",
  color = "grey30"
) +
labs(
  title = "Figura 3.38: Distribuição da variável `ses`"
) +
tema

```



Temos uma leve assimetria à direita. O Coeficiente de Assimetria de Fisher-Pearson Ajustado é 0.9535379. Vamos verificar se podemos usar uma transformação nos dados.

```

dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[[i]] <- antracose$ses %>%
    boxcox(p[i]) %>%
    calcular_metrica()
}

```

```
## Warning in log(x): NaNs produzidos
```

```
p[min_rank(dps)[1]]
```

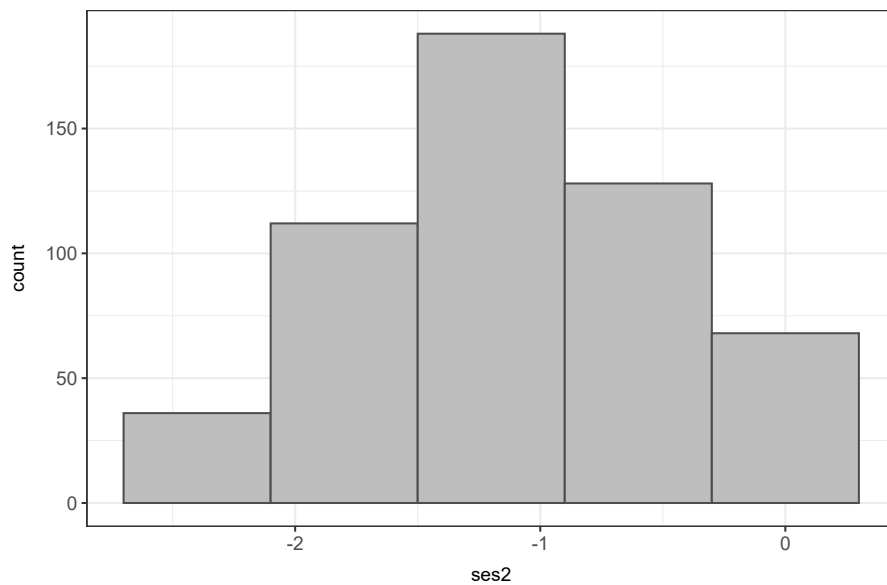
```
## [1] 0.4
```

O valor $p = 0.4$ minimiza a nossa medida de avaliação. Vamos agora plotar os dados transformados e avaliar novamente o coeficiente de assimetria.

```
antracose$ses2 <- boxcox(antracose$ses, p[min_rank(dps)[1]])  
  
antracose %>%  
  ggplot(aes(x = ses2)) +  
    geom_histogram(  
      binwidth = .6,  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.39: Distribuição da variável `antracose`"  
    ) +  
    tema
```

```
## Warning: Removed 1920 rows containing non-finite values (`stat_bin()`).
```

Figura 3.39: Distribuição da variável `antracose`



Notamos que o histograma se tornou um pouco mais simétrico. Vamos avaliar os valores dos coeficientes de assimetria.

```
e1071::skewness(antracose$ses, na.rm = TRUE, type = 2)
```

```
## [1] 0.9793336
```

```
e1071::skewness(antracose$ses2, na.rm = TRUE, type = 2)
```

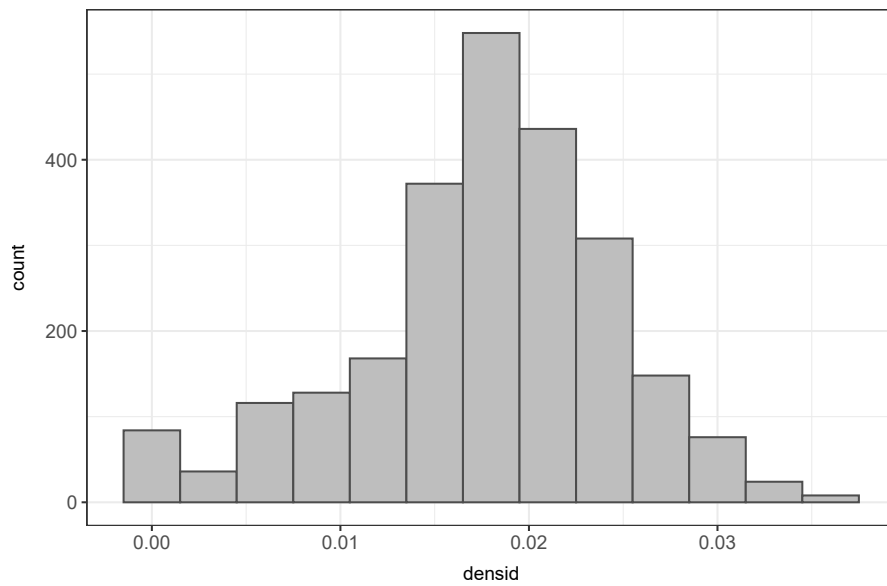
```
## [1] -0.1564781
```

De fato, a transformação da variável nos trouxe uma melhoria significativa na simetria dos dados.

Para a variável `densid`:

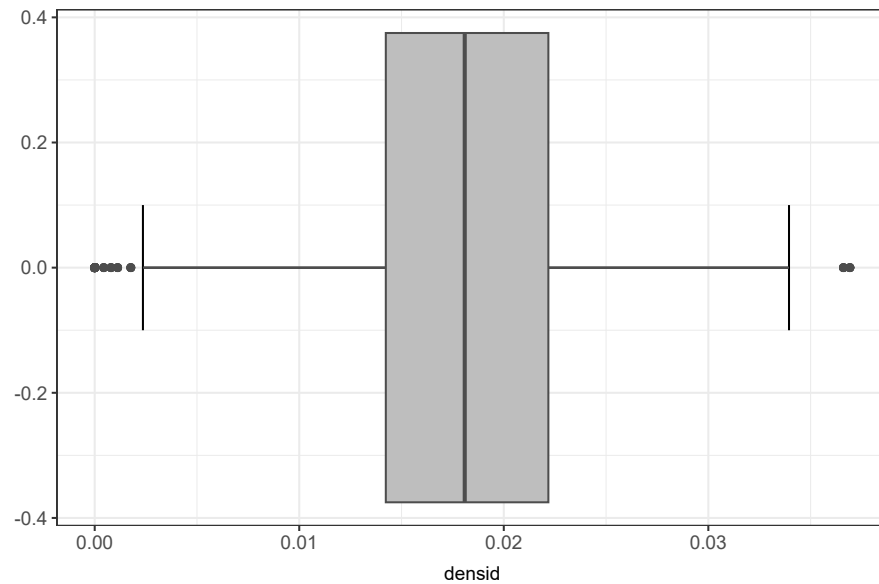
```
antracose %>%  
  ggplot(aes(x = densid)) +  
    geom_histogram(  
      binwidth = .003,  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.40: Distribuição da variável `densid`"  
    ) +  
    tema
```


Figura 3.40: Distribuição da variável `densid`



```
antracose %>%  
  ggplot(aes(x = densid)) +  
    stat_boxplot(geom = "errorbar", width = 0.2) +  
    geom_boxplot(  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.41: Distribuição da variável `densid`"  
    ) +  
    tema
```

Figura 3.41: Distribuição da variável `densid`



A distribuição já nos parece bastante adequada, com coeficiente de assimetria - 0.4460641.

```
dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[[i]] <- antracose$densid %>%
    boxcox(p[i]) %>%
    calcular_metrica()
}

p[min_rank(dps)[1]]
```

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

Para os diversos valores de p , o que minimiza a nossa métrica é justamente $p = 1$, que nos dá a transformação identidade. Em outras palavras, as transformações com as quais estamos trabalhando, não nos darão melhores resultados.

Por fim, vamos à variável `distmin`:

```
antracose %>%
  ggplot(aes(x = distmin)) +
```

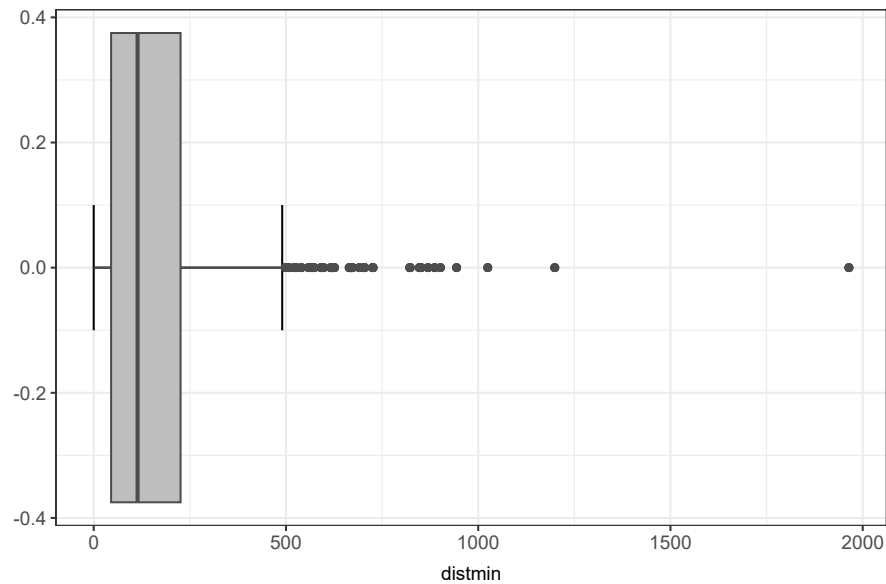
```
geom_histogram(  
  binwidth = 190,  
  fill = "grey",  
  color = "grey30"  
) +  
labs(  
  title = "Figura 3.42: Distribuição da variável `distmin`"  
) +  
tema
```

Figura 3.42: Distribuição da variável `distmin`



```
antracose %>%  
  ggplot(aes(x = distmin)) +  
  stat_boxplot(geom = "errorbar", width = 0.2) +  
  geom_boxplot(  
    fill = "grey",  
    color = "grey30"  
  ) +  
  labs(  
    title = "Figura 3.43: Distribuição da variável `distmin`"  
  ) +  
  tema
```

Figura 3.43: Distribuição da variável `distmin`



Temos uma forte assimetria à direita. O Coeficiente de Assimetria de Fisher-Pearson Ajustado é 2.9669771. Vamos verificar se podemos usar uma transformação nos dados.

```
dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[[i]] <- antracose$distmin %>%
    boxcox(p[i]) %>%
    calcular_metrica()
}

p[min_rank(dps)[1]]
```

```
## [1] 0.2
```

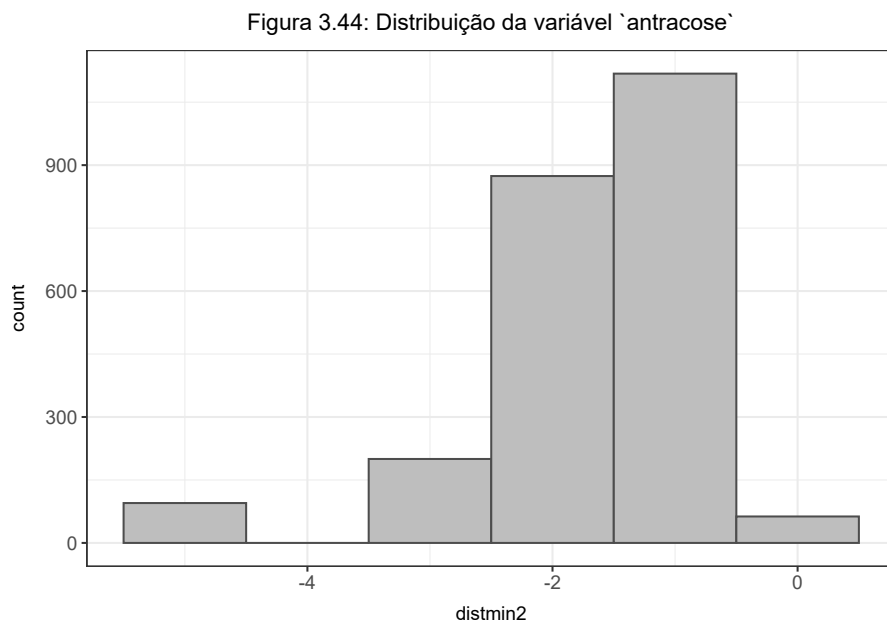
O valor $p = 0.2$ minimiza a nossa medida de avaliação. Vamos agora plotar os dados transformados e avaliar novamente o coeficiente de assimetria.

```
antracose$distmin2 <- boxcox(antracose$antracose, p[min_rank(dps)[1]])

antracose %>%
  ggplot(aes(x = distmin2)) +
```

```
geom_histogram(  
  binwidth = 1,  
  fill = "grey",  
  color = "grey30"  
) +  
labs(  
  title = "Figura 3.44: Distribuição da variável `antracose`"  
) +  
tema
```

```
## Warning: Removed 102 rows containing non-finite values (`stat_bin()`).
```



Notamos uma pequena melhora, porém ainda há uma concentração de dados em torno de um valor inicial. Analisemos o coef. de assimetria antes e depois da transformação.

```
e1071::skewness(antracose$distmin, na.rm = TRUE, type = 2)
```

```
## [1] 2.966977
```

```
e1071::skewness(antracose$distmin2, na.rm = TRUE, type = 2)
```

```
## [1] -1.819497
```

Exercício 3.3

Considere as variáveis `Peso` e `Altura` de homens do conjunto de dados `rehabcardio`. Determine o número de classes para os histogramas correspondentes por meio de (3.26) e (3.27) e construa-os.

Solução. Inicialmente vamos filtrar o conjunto de dados original selecionando apenas os indivíduos do sexo masculino.

```
rehabcardio_h <- rehabcardio %>%  
  filter(Genero == "M")
```

Vamos agora definir as funções `freedman_diaconis` e `sturges` para calcular a amplitude das classes.

```
freedman_diaconis <- function(x) {  
  s <- sd(x, na.rm = TRUE)  
  n <- length(x)  
  
  1.349 * s * ((log(n) / n)^(1/3))  
}  
  
sturges <- function(x) {  
  r <- range(rehabcardio_h$Peso, na.rm = T)  
  w = r[2] - r[1]  
  n <- length(x)  
  
  w / (1 + 3.322 * log(n))  
}
```

Com as funções definidas, podemos construir os histogramas usando essa amplitude de classes.

```
rehabcardio_h %>%  
  ggplot(aes(Peso)) +  
    geom_histogram(  
      binwidth = freedman_diaconis(rehabcardio_h$Peso),  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.45: Distribuição do peso utilizando Freedman-Diaconis"  
    ) +  
    tema
```

```
## Warning: Removed 5 rows containing non-finite values (`stat_bin()`).
```

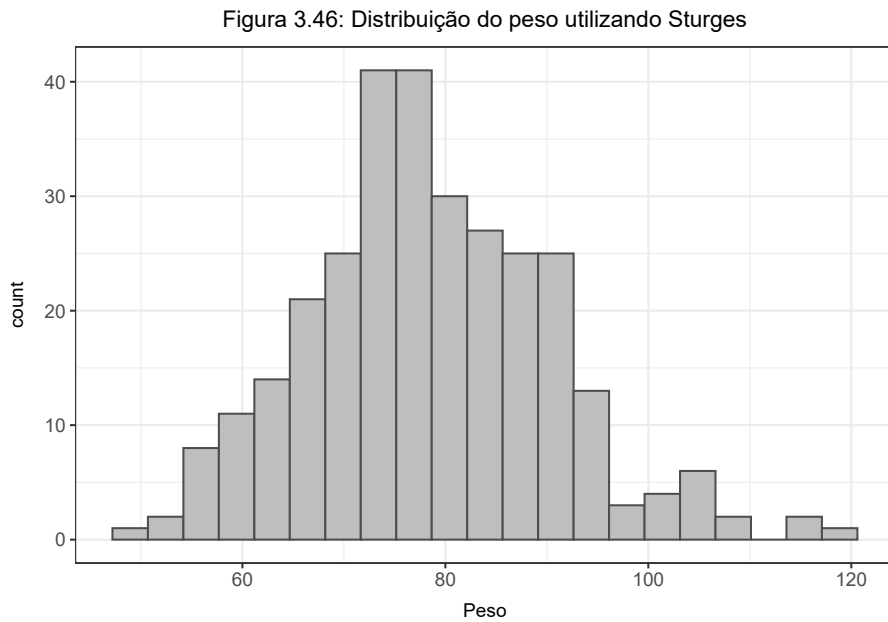
Figura 3.45: Distribuição do peso utilizando Freedman-Diaconis



```
rehabcardio_h %>%  
  ggplot(aes(Peso)) +  
    geom_histogram(  
      binwidth = sturges(rehabcardio_h$Peso),  
      fill = "grey",  
      color = "grey30"  
    ) +
```

```
labs(
  title = "Figura 3.46: Distribuição do peso utilizando Sturges"
) +
tema
```

```
## Warning: Removed 5 rows containing non-finite values (`stat_bin()`).
```



```
rehabcardio_h %>%
  ggplot(aes(Altura)) +
  geom_histogram(
    binwidth = freedman_diaconis(rehabcardio_h$Altura),
    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = "Figura 3.47: Distribuição da altura utilizando Freedman-Diaconis"
  ) +
tema
```

```
## Warning: Removed 6 rows containing non-finite values (`stat_bin()`).
```


Figura 3.47: Distribuição da altura utilizando Freedman-Diaconis



```
rehabcardio_h %>%  
  ggplot(aes(Altura)) +  
    geom_histogram(  
      binwidth = sturges(rehabcardio_h$Altura),  
      fill = "grey",  
      color = "grey30"  
    ) +  
    labs(  
      title = "Figura 3.48: Distribuição da altura utilizando Sturges"  
    ) +  
    tema
```

```
## Warning: Removed 6 rows containing non-finite values (`stat_bin()`).
```

Figura 3.48: Distribuição da altura utilizando Sturges

**Exercício 3.4**

Considere o arquivo `vento.xls`. Observe o valor atípico 61, 1, que na realidade ocorreu devido a uma forte tempestade no dia 2 de dezembro. Calcule as medidas de posição e dispersão apresentadas na Seção 3.3. Quantifique o efeito do valor atípico indicado nessas medidas.

Solução. x

Exercício 3.5

Construa gráficos ramo-e-folhas e boxplot para os dados do Exercício 4.

Solução. x

Exercício 3.6

Transforme os dados do Exercício 4 por meio de (3.23) com $p = 0, 1/4, 1/3, 1/2, 3/4$ e escolha a melhor alternativa de acordo com a medida d_p , dada em (3.24).

Solução. x

Exercício 3.7

Analise a variável Temperatura do arquivo `poluicao.xls`.

Solução. [x](#)

Exercício 3.8

Analise a variável Salário de administradores, disponível no arquivo `salarios.xls`.

Solução. [x](#)

Exercício 3.9

Construa um gráfico ramo-e-folhas e um boxplot para os dados de precipitação atmosférica de Fortaleza disponíveis no arquivo `precipitacao.manchas solares` disponíveis no arquivo `manchas.xls`.

Solução. [x](#)

Exercício 3.10

Construa gráficos de quantis e de simetria para os dados de arquivo manchas solares disponíveis no arquivo `manchas.xls`.

Solução. [x](#)

Exercício 3.11

Uma outra medida de assimetria é

$$A = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1},$$

que é igual a zero no caso de distribuições simétricas. Calcule-a para os dados do Exercício 4.

Solução. [x](#)

Exercício 3.12

Os dados disponíveis no arquivo `endometrone.xls` são provenientes de um estudo em que o objetivo é verificar se existe diferença entre os grupos de doentes e controles quanto a algumas características observadas.

- O pesquisador responsável pelo estudo tem a seguinte pergunta: pacientes doentes apresentam mais dor na menstruação do que as pacientes não doentes? Que tipo de análise você faria para responder essa pergunta utilizando as técnicas discutidas neste capítulo? Faça-a e tire suas conclusões.
- Compare as distribuições das variáveis idade e concentração de PCR durante a menstruação (PCR_a) para pacientes dos grupos controle e doente utilizando medidas resumo (mínimo, máximo, quartis, mediana, média, desvio padrão, etc.), *boxplots*, histogramas, gráficos de médias e gráficos QQ. Como você considerou os valores $< 0,5$ da variável PCR_a nesses cálculos? Você sugeriria uma outra maneira para considerar tais valores?
- Compare a distribuição da variável número de gestações para os dois grupos por intermédio de uma tabela de frequências. Utilize um método gráfico para representá-la.

Solução. x

Exercício 3.13

Os dados encontrados no arquivo `enforco.xls` são provenientes de um estudo sobre teste de esforço cardiopulmonar em pacientes com insuficiência cardíaca. As variáveis medidas durante a realização do teste foram observadas em quatro momentos distintos: repouso (REP), limiar anaeróbico (LAN), ponto de compensação respiratório (PCR) e pico (PICO). As demais variáveis são referentes às características demográficas e clínicas dos pacientes e foram registradas uma única vez.

- Descreva a distribuição da variável consumo de oxigênio (v_{O_2}) em cada um dos quatro momentos de avaliação utilizando medidas resumo (mínimo, máximo, quartis, mediana, média, desvio padrão, etc.), *boxplots* e histogramas. Você identifica algum paciente com valores de consumo de oxigênio discrepantes? Interprete os resultados.
- Descreva a distribuição da classe funcional NYHA por meio de uma tabela de frequências. Utilize um método gráfico para representar essa tabela.

Solução. x

Exercício 3.14

Na tabela abaixo estão indicadas as durações de 335 lâmpadas.

Duração (horas)	Número de lâmpadas
0 – 100	82
100 – 200	71

Duração (horas)	Número de lâmpadas
200 – 300	68
300 – 400	56
400 – 500	43
500 – 800	15

- Esboce o histograma correspondente.
- Calcule os quantis de ordem $p = 0, 1, 0, 3, 0, 5, 0, 7$ e $0, 9$.

Solução. x

Exercício 3.15

Os dados apresentados na tabela abaixo referem-se aos instantes nos quais o centro de controle operacional de estradas rodoviárias recebeu chamados solicitando algum tipo de auxílio em duas estradas num determinado dia.

Estrada 1	12:07:00 AM	15:58:00 AM	01:24:00 AM	01:35:00 AM	02:05:00 AM
	03:14:00 AM	03:25:00 AM	03:46:00 AM	05:44:00 AM	05:56:00 AM
	06:36:00 AM	07:26:00 AM	07:48:00 AM	09:13:00 AM	12:05:00 PM
	12:48:00 PM	01:21:00 PM	02:22:00 PM	05:30:00 PM	06:00:00 PM
	07:53:00 PM	09:15:00 PM	09:49:00 PM	09:59:00 PM	10:53:00 PM
	11:27:00 PM	11:49:00 PM	11:57:00 PM		
Estrada 2	12:03:00 AM	01:18:00 AM	04:35:00 AM	06:13:00 AM	06:59:00 AM
	08:03:00 AM	10:07:00 AM	12:24:00 PM	01:45:00 PM	02:07:00 PM
	03:23:00 PM	06:34:00 PM	07:19:00 PM	09:44:00 PM	10:27:00 PM
	10:52:00 PM	11:19:00 PM	11:29:00 PM	11:44:00 PM	

- Construa um histograma para a distribuição de frequências dos instantes de chamados em cada uma das estradas.
- Calcule os intervalos de tempo entre as sucessivas chamadas e descreva-os, para cada uma das estradas, utilizando medidas resumo e gráficos do tipo *boxplot*. Existe alguma relação entre o tipo de estrada e o intervalo de tempo entre as chamadas?

- c) Por intermédio de um gráfico do tipo QQ, verifique se a distribuição da variável Intervalo de tempo entre as chamadas em cada estrada é compatível com um modelo normal. Faça o mesmo para um modelo exponencial. Compare as distribuições de frequências correspondentes às duas estradas.

Solução. x

Exercício 3.16

As notas finais de um curso de Estatística foram: 7, 5, 4, 5, 6, 3, 8, 4, 5, 4, 6, 4, 5, 6, 4, 6, 6, 3, 8, 4, 5, 4, 5, 5 e 6.

- Calcule a mediana, os quartis e a média.
- Separe o conjunto de dados em dois grupos denominados **aprovados**, com nota pelo menos igual a 5, e **reprovados**, com notas menores do que 5. Compare a variância das notas desses dois grupos.

Solução. x

Exercício 3.17

Considere o seguinte resumo descritivo da pulsação de estudantes com atividade física intensa e fraca:

Atividade	N	Media	Mediana	DP	Min	Max	Q1	Q3
Intensa	30	79,6	82	10,5	62	90	70	85
Fraca	30	73,1	70	9,6	58	92	63	77

Indique se as seguintes afirmações estão corretas, justificando a sua respostas:

- 5% e 50% dos estudantes com atividade física intensa e fraca, respectivamente, tiveram pulsação inferior a 70.
- A proporção de estudantes com fraca atividade física com pulsação inferior a 63 é menor que a proporção de estudantes com atividade física intensa com pulsação inferior a 70.
- A atividade física não tem efeito na média da pulsação dos estudantes.
- Mais da metade dos estudantes com atividade física intensa têm pulsação maior que 82.

Solução. x

Exercício 3.18

Considere os gráficos *boxplot* da Figura 3.34. Quais deles correspondem às pulsações dos estudantes submetidos a atividade física intensa e fraca?

- a) A e B
- b) B e D
- c) A e C
- d) B e C

Solução. x

Exercício 3.19

Os histogramas apresentados na Figura 3.35 mostram a distribuição das temperaturas ($^{\circ}\text{C}$) ao longo de vários dias de investigação para duas regiões (R1 e R2). Indique se as afirmações abaixo estão corretas, justificando as respostas:

- a) As temperaturas das regiões R1 e R2 têm mesma média e mesma variância.
- b) Não é possível comparar as variâncias.
- c) A temperatura média da região R2 é maior que a de R1.
- d) As temperaturas das regiões R1 e R2 têm mesma média e variância diferentes.

Solução. x

Exercício 3.20

Na companhia A, a média dos salários é 10000 unidades e o 3 quartil é 5000. Responda as seguintes perguntas, justificando a sua respostas:

- a) Se você se apresentasse como candidato a funcionário nessa firma e se o seu salário fosse escolhido ao acaso entre todos os possíveis salários, o que seria mais provável: ganhar mais ou menos que 5000 unidades?
- b) Suponha que na companhia B a média dos salários seja 7000 unidades, a variância praticamente zero e o salário também seja escolhido ao acaso. Em qual companhia você se apresentaria para procurar emprego, com base somente no salário?

Solução. x

Exercício 3.21

Num conjunto de dados, o primeiro quartil é 10, a mediana é 15 e o terceiro quartil é 20. Indique quais das seguintes afirmativas são verdadeiras, justificando sua resposta:

- a) A distância interquartis é 5.
- b) O valor 32 seria considerado outlier segundo o critério utilizado na construção do boxplot.
- c) A mediana ficaria alterada de 2 unidades se um ponto com valor acima do terceiro quartil fosse substituído por outro 2 vezes maior.
- d) O valor mínimo é maior do que zero.

Solução. x

Exercício 3.22

A bula de um medicamento A para dor de cabeça afirma que o tempo médio para que a droga faça efeito é de 60 seg com desvio padrão de 10 seg. A bula de um segundo medicamento B afirma que a média correspondente é de 60 seg com desvio padrão de 30 seg. Sabe-se que as distribuições são simétricas. Indique quais das seguintes afirmativas são verdadeiras, justificando sua resposta:

- a) Os medicamentos são totalmente equivalentes com relação ao tempo para efeito pois as médias são iguais.
- b) Com o medicamento A, a probabilidade de cura de sua dor de cabeça antes de 40 seg é maior do que com o medicamento B.
- c) Com o medicamento B, a probabilidade de você ter sua dor de cabeça curada antes de 60 seg é maior que com o medicamento A.

Solução. x

Exercício 3.23

A tabela abaixo representa a distribuição do número de dependentes por empregado de uma determinada empresa.

Dependentes	Frequência
1	40
2	50
3	30
4	20
5	10

Dependentes	Frequência
Total	150

A mediana, média e moda cujos valores calculados por quatro estagiários, foram:

- a) 50; 15; 50
- b) 1; 2,1; 1
- c) 50,5; 50; 50
- d) 1; 1; 1

Indique qual deles está correto, justificando sua resposta.

Solução. x

Exercício 3.24

Com relação ao Exercício 23, qual a porcentagem de empregados da empresa com 2 ou mais dependentes?

- a) 40,1%
- b) 50,1%
- c) 60,3%
- d) 73,3%

Solução. x

Exercício 3.25

Num estudo na área de Oncologia, o número de vasos que alimentam o tumor está resumido na seguinte tabela.

Número de vasos	Frequência
0 – 5	8 (12%)
5 – 10	23 (35%)
10 – 15	12 (18%)
15 – 20	9 (14%)
20 – 25	8 (12%)
25 – 30	6 (9%)
Total	66 (100%)

Indique a resposta correta.

- a) O primeiro quartil é 25%.
- b) A mediana está entre 10 e 15.
- c) O percentil de ordem 10% é 10.
- d) A distância interquartis é 50.
- e) Nenhuma das respostas anteriores.

Solução. x

Exercício 3.26

Utilizando o mesmo enunciado da questão anterior, indique a resposta correta:

- a) Não é possível estimar nem a média nem a variância com esses dados.
- b) A variância é menor que 30.
- c) A média estimada é 12,8.
- d) Em apenas 35% dos casos, o número de vasos é maior que 10.
- e) Nenhuma das anteriores.

Solução. x

Exercício 3.27

Em dois estudos realizados com o objetivo de estimar o nível médio de colesterol total para uma população de indivíduos saudáveis observaram-se os dados indicados na tabela seguinte:

Tabela 3.11: Tabela A: Medidas descritivas dos estudos A e B

Estudo	n	Média	Desvio padrão
A	100	160 mg/dL	60 mg/dL
B	49	150 mg/dL	35 mg/dL

Indique a resposta correta:

- a) Não é possível estimar o nível médio de colesterol populacional só com esses dados.
- b) Se os dois estudos foram realizados com amostras da mesma população não deveria haver diferença entre os desvios padrões amostrais.
- c) Com os dados do estudo B, o colesterol médio populacional pode ser estimado com mais precisão do que com os dados do estudo A.
- d) Ambos os estudos sugerem que a distribuição do colesterol na população é simétrica.

- e) Nenhuma das respostas anteriores.

Solução. x

Exercício 3.28

Considere um conjunto de dados $\{X_1, \dots, X_n\}$

- a) Obtenha a média e a variância de W_1, \dots, W_n em que $W_l = X_l + k$ com k denotando uma constante, em termos da média e da variância de X .
- b) Calcule a média e a variância de v_1, \dots, V_n em que $V_i = kX_i$ com k denotando uma constante, em termos da média e da variância de X .

Solução. x

Exercício 3.29

Prove que S^2 dado por (3.10) é um estimador não enviesado da variância populacional.

Solução. x

Exercício 3.30

Considere os valores X_1, \dots, X_n de uma variável X , com média \bar{X} e desvio padrão S . Mostre que a variável Z , cujos valores são $Z_i = (X_i - \bar{X})/S$, $i = 1, \dots, n$, tem média 0 e desvio padrão 1.

Solução. x

Exercício 3.31

Prove a relação (3.8). Como ficaria essa expressão para S^2 ?

Solução. x

Exercício 3.32

Considere uma amostra aleatória simples X_1, \dots, X_n de uma variável X que assume o valor 1 com probabilidade $0 < p < 1$ e o valor 0 com probabilidade $1 - p$. Seja $\hat{p} = n^{-1} \sum_{i=1}^n X_i$. Mostre que:

- i) $E(X_i) = p$ e $Var(X_i) = p(1 - p)$.

- ii) $E(\hat{p}) = p$ e $Var(\hat{p}) = p(1 - p)/n$.
- iii) $0 < Var(X_i) < 0,25$.

Com base nesses resultados, utilize o Teorema Limite Central [ver Sen et al. (2009), por exemplo] para construir um intervalo de confiança aproximado conservador (ie, com a maior amplitude possível) para p . Utilize o Teorema de Sverdrup [ver Sen et al. (2009), por exemplo] para construir um intervalo de confiança aproximado para p com amplitude menor que a do intervalo mencionado acima. Veja também, Bussab e Morettin (2017).

Solução. x

Exercício 3.33

Com a finalidade de entender a diferença entre “desvio padrão” e “erro padrão”,

- a) Simule 10000 dados de uma distribuição normal com média 12 e desvio padrão 4. Construa o histograma correspondente, calcule a média e o desvio padrão amostrais e compare os valores obtidos com aqueles utilizados na geração dos dados.
- b) Simule 500 amostras de tamanho $n = 4$ dessa população. Calcule a média amostral de cada amostra, construa o histograma dessas médias e estime o correspondente desvio padrão (que é o erro padrão da média).
- c) Repita os passos a) e b) com amostras de tamanhos $n = 9$ e $n = 100$. Comente os resultados comparando-os com aqueles preconizados pela teoria.
- d) Repita os passos a) - c) simulando amostras de uma distribuição qui-quadrado com 3 graus de liberdade.

Solução. x

4

Análise de dados de duas variáveis

4.1 Introdução

4.2 Duas variáveis qualitativas

4.3 Duas variáveis quantitativas

4.4 Uma variável qualitativa e outra quantitativa

4.5 Notas de capítulo

4.6 Exercícios



5

Análise de dados de várias variáveis

5.1 Introdução

5.2 Gráficos para três variáveis

5.3 Gráficos para quatro ou mais variáveis

5.4 Medidas resumo multivariadas

5.5 Tabelas de contingência de múltiplas entradas

5.6 Notas de capítulo

5.7 Exercícios



6

Análise de Regressão

6.1 Introdução

6.2 Regressão linear simples

6.3 Regressão linear múltipla

6.4 Regressão para dados longitudinais

6.5 Regressão logística

6.6 Notas de capítulo

6.7 Exercícios



7

Análise de Sobrevivência

7.1 Introdução

7.2 Estimação da função de sobrevivência

7.3 Comparação de curvas de sobrevivência

7.4 Regressão para dados de sobrevivência

7.5 Notas de capítulo

7.6 Exercícios



Parte II

Aprendizado Supervisionado



8

Regularização e Modelos Aditivos Generalizados

8.1 Introdução

8.2 Regularização

8.3 Modelos aditivos generalizados (GAM)

8.4 Notas de capítulo

8.5 Exercícios



9

Classificação por meio de técnicas clássicas

9.1 Introdução

9.2 Classificação por regressão logística

9.3 Análise discriminante linear

9.4 Classificador do vizinho mais próximo

9.5 Algumas extensões

9.6 Notas de capítulo

9.7 Exercícios



10

Algoritmos de Suporte Vetorial

10.1 Introdução

10.2 Fundamentação dos algoritmos de suporte vetorial

10.3 Classificador de margem máxima

10.4 Classificador de margem flexível

10.5 Classificador de margem não linear

10.6 Regressão por algoritmos de suporte vetorial

10.7 Notas de capítulo

10.8 Exercícios



11

Árvores e Florestas

11.1 Introdução

11.2 Classificação por árvores

11.3 *Bagging, boosting* e florestas

11.4 Árvores para regressão

11.5 Notas de capítulo

11.6 Exercícios



12

Redes neurais

12.1 Introdução

12.2 *Perceptron*

12.3 Redes com camadas ocultas

12.4 O algoritmo de retropropagação (*backpropagation*)

12.5 Aprendizado profundo (*Deep learning*)

12.6 Notas de capítulo

12.7 Exercícios



Parte III

Aprendizado não Supervisionado



13

Análise de Agrupamentos

13.1 Introdução

13.2 Estratégias de agrupamento

13.3 Algoritmos hierárquicos

13.4 Algoritmos de partição: K-médias

13.5 Notas de capítulo

13.6 Exercícios



14

Redução de dimensionalidade

14.1 Introdução

14.2 Análise de Componentes Principais

14.3 Análise fatorial

14.4 Análise de componentes independentes

14.5 Notas de capítulo

14.6 Exercícios



A

Otimização numérica

A.1 Introdução

A.2 O método de Newton-Raphson

A.3 O método scoring

A.4 O método de Gauss-Newton

A.5 Métodos Quase-Newton

A.6 Aspectos computacionais

A.7 Notas de capítulo

A.8 Exercícios



B

Noções de simulação

B.1 Introdução

B.2 Método Monte Carlo

B.3 Simulação de variáveis discretas

B.4 Simulação de variáveis contínuas

B.5 Simulação de vetores aleatórios

B.6 Métodos de reamostragem

B.7 Notas de capítulo

B.8 Exercícios



C

Algoritmos para dados aumentados

C.1 Introdução

C.2 O algoritmo EM

C.3 O algoritmo EM Monte Carlo

C.4 Cálculo de erros padrões

C.5 O algoritmo para dados aumentados

C.6 Exercícios



Bibliografia

Tales de Carvalho, Ana Luiza Hallal Curi, Dalton Francisco Andrade, Julio da Motta Singer, Magnus Benetti, and Alfredo José Mansur. Reabilitação cardiovascular de portadores de cardiopatia isquêmica submetidos a tratamento clínico, angioplastia coronariana transluminal percutânea e revascularização cirúrgica do miocárdio. *Arquivos Brasileiros de Cardiologia*, 88(1):72–78, Jan 2007. ISSN 0066-782X. doi: 10.1590/S0066-782X2007000100012. URL <https://doi.org/10.1590/S0066-782X2007000100012>.

Pedro Alberto Morettin and Julio da Motta Singer. *Estatística e Ciência de Dados*. LTC, Rio de Janeiro, 2022.