Jeidsan A. da C. Pereira

Estatística e Ciência de Dados

Notas e solução dos exercícios

Conteúdo

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, $\overline{x}(\alpha)$, 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 capitulo. 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].

Funções auxiliares

Durante a solução dos exercícios, repetiremos muitas vezes algumas ações (como o carregamento de um arquivo, o cálculo de medidas de resumo para uma determinada variável, etc). Para evitar replicar o código para essas ações, vaos criar aqui algumas funções.

A primeira delas visa carregar um arquivo de dados:

```
statds_read <- function(file, type = "csv", ...) {
  file <- paste0(data_dir, file)

  switch(
    type,
    "csv" = readr::read_csv(file, ...),
    "xls" = readxl::read_xls(file, ...)
  )
}
```

A segunda função tem o objetivo de gravar os dados ajustados em um arquivo:

```
statds_write <- function(data, file, ...) {
  file <- paste0(data_dir, file)
  readr::write_csv(data, file, ...)
}
```

A nossa proxima função servirá para calcular um resumo de uma coluna de um data frame:

```
statds_summarise <- function(data, x, ...) {
  x <- enquo(x)
  data %>%
    group_by(...) %>%
    summarize(
        `n` = n(),
        `Média` = mean(!!x, na.rm = TRUE),
        `Variância` = var(!!x, na.rm = TRUE),
        `Desvio Padrão` = sd(!!x, na.rm = TRUE),
        `Min.` = min(!!x),
        `Q1` = quantile(!!x, c(0.25), na.rm = TRUE)[[1]],
        `Mediana` = median(!!x, na.rm = TRUE),
        `Q3` = quantile(!!x, c(0.75), na.rm = TRUE)[[1]],
        `Máx.` = max(!!x),
        `IQR` = IQR(!!x, na.rm = TRUE)
    )
}
```

A próxima função tem o objetivo de construir a tabela de contingência para duas variáveis categóricas:

```
statds_crosstable <- function(data, x, y, totals = TRUE) {
  x <- enquo(x)
  y <- enquo(y)

  data %>%
    group_by(!!x, !!y) %>%
    count() %>%
    bind_rows(
      group_by(., !!x) %>%
        summarise(n = sum(n)) %>%
        mutate(regiao = 'Total')
    ) %>%
    bind_rows(
      group_by(., !!y) %>%
        summarise(n = sum(n)) %>%
        mutate(estado = 'Total')
    ) %>%
    spread(!!y, n)
}
```

A função a seguir calcula o coeficiente de determinação (ou o ganho de varância) para uma variável quantitativa em relação a outra qualitativa.

```r
statds_rsquared <- function(data, x, y) {
  x <- enquo(x)
  y <- enquo(y)

  total <- data %>%
              select(!!x)

  var_total <- var(total)[[1]]

  groups <- data %>%
              group_by(!!y) %>%
              summarise(
                var = var(!!x, na.rm = TRUE)
              )

  var_groups <- mean(groups$var)

  1 - (var_groups / var_total)
}
```

Pendências

- Ajustar títulos do gráfico colocando a parte "Figura x.y:" em negrito;
- Ajustar capítulos 2 e 3 pra usarem as funções `statds_()` definidas acima;
- Melhorar os totais da função `statds_crosstable()` e utilizá-la nos exercícios;
- Exercício 2.2;
- Revisar estrutura do Exercício 3.5;
- Complementar a análise do Exercício 3.12;
- Reavaliar Exercício 3.17 - Questão mal formulada;
- Reavaliar Exercício 3.18 - Alternativas não parecem conter a resposta correta;
- Reavaliar Exercício 3.27;
- Exercício 3.29;
- Exercício 3.31;
- Exercício 3.32;
- Exercício 3.33;
- Validar cálculo de R^2 no Exercício 4.1;
- Revisar cálculo de conceito de R^2 no último item do Exercício;
- Reavaliar e interpretar resultados do Exercício 4.9;

1

Estatística, Ciência de Dados e Megadados

1.1 Introdução

Atualmente, os termos *Data Science* (**Ciência de Dados**) e *Big Data* (**Megadadas**) 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 consi- derado na análise de regressão usual, em que se pretende entender como cada variável preditora Ro está associada com a variável resposta. O objetivo do aprendizado automático é selecionar o mo- delo 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 [Morettin 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 conclusães 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.

a) Com base nessa planilha, apresente um dicionário com a especificação das variáveis segundo as indicações da Seção 2.2 e construa a planilha correspondente.
b) 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: 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 nausea 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: Tabela 2.2: Distribuição dos pacientes conforme o sexo

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

[1]Trata-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: 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: 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: 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: 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: 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 - Interior2 - Capital 3 - Outra

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

Tabela 2.8: 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: 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: 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 cardiaca` e `VO2` no repouso além do quociente `VE/VCO2`, as correspondentes porcentagens relativamente ao máximo, o quociente `VO2/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 estatistica. 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) a definição das variáveis.
b) 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 `cidades.xls` contém informações demográficas de 3554 municípios brasileiros.

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

a) Importe-a para permitir a análise por meio do software R, indicando os problemas encontrados nesse processo além de sua solução.
b) Use o comando `summary` para obter um resumo das variáveis do arquivo.
c) 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.

```
cidades <- read_csv(paste0(data_dir, "cidades.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 pacore **readxl**.

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

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

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

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

```
colnames(cidades) <- str_to_lower(colnames(cidades))
```

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(cidades)
```

```
##     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          11oumais
##  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 `cidades.xls`

Variável	Tipo	Número de observações faltantes
munic	Alfanumérica	0
uf	Alfanumérica	0
codigo	Numérica	0
poptot	Numérica	0
cres_pop	Numérica	0
popurb	Numérica	0
pibtot	Numérica	14
cres_pib	Numérica	14
grau1	Numérica	13
grau2	Numérica	13
superior	Numérica	13
11oumais	Numérica	13
empregad	Numérica	14
microemp	Numérica	14
peqemp	Numérica	14
medemp	Numérica	14
graemp	Numérica	14

Exercício 2.9

Preencha a ficha de inscrição do Centro de Estatistica Aplicada www.ime.usp.br/-cea[3] 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, $\overline{x}(\alpha)$, 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 $X = \{x_1, \cdots, x_n\}$ com $n \in \mathbb{N}$, o k-ésimo **momento centrado** de X é dado por [Morettin and Singer, 2022, p. 50]:

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

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 consequente- mente, 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

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

[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 $\overline{X}$ 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

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

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

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

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

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

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

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

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

[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 cardiacos. 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,

a) Construa distribuições de frequências para as variáveis qualitativas.
b) Construa histogramas, *boxplots* e gráficos de simetria para as variáveis continuas.
c) Construa uma tabela com medidas resumo para as variáveis continuas.
d) Avalie a compatibilidade de distribuições normais para as variáveis continuas 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 vatiáveis em `rehabcardio.xls`",
  label = NA,
  digits = 2,
  align = "c",
  format.args = list(decimal.mark = ",")
)
```

Tabela 3.1: Tabela 3.1: Tipos de dados das vatiáveis em `rehabcardio.xls`

Registro	double
Genero	character
Origem	character
Grupo	character
Nascimento	double
Coleta	double
Programa	double
Lesoes	double
IAM1	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),
                    Diabete = as.logical(Diabete),
                    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 tabela 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: 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: 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

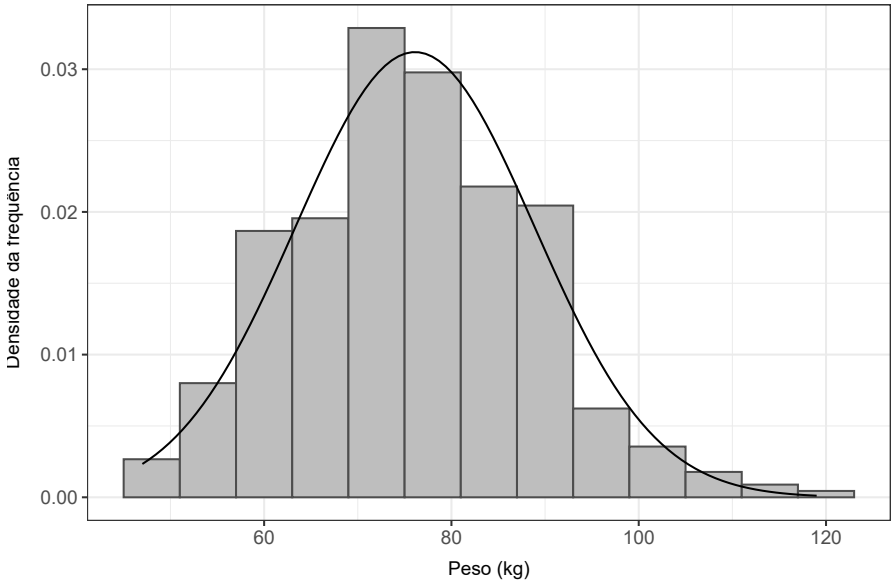
Histogramas, *boxplots* e gráficos de simetria para as variáveis continuas

```
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()`).
```

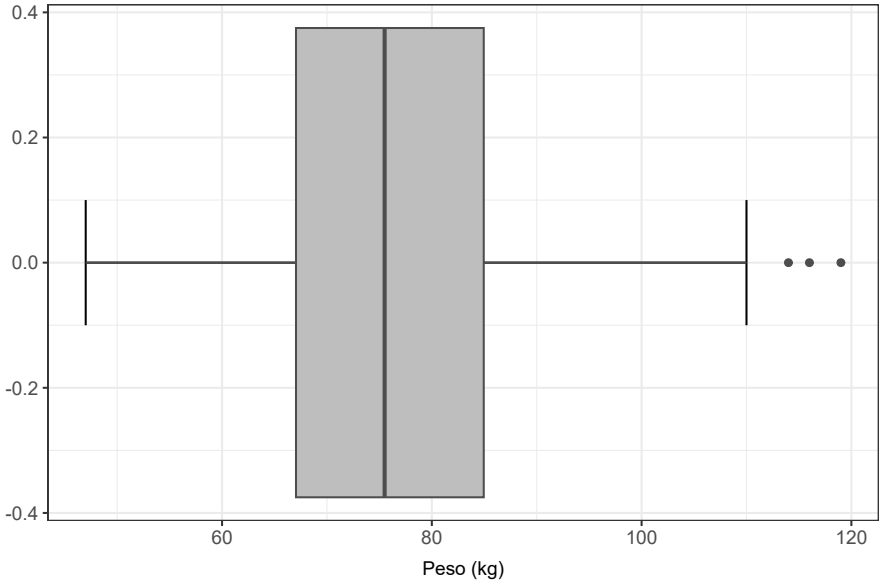
Figura 3.1: Distribuição de frequência dos pesos dos pacientes




```
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()`).
```

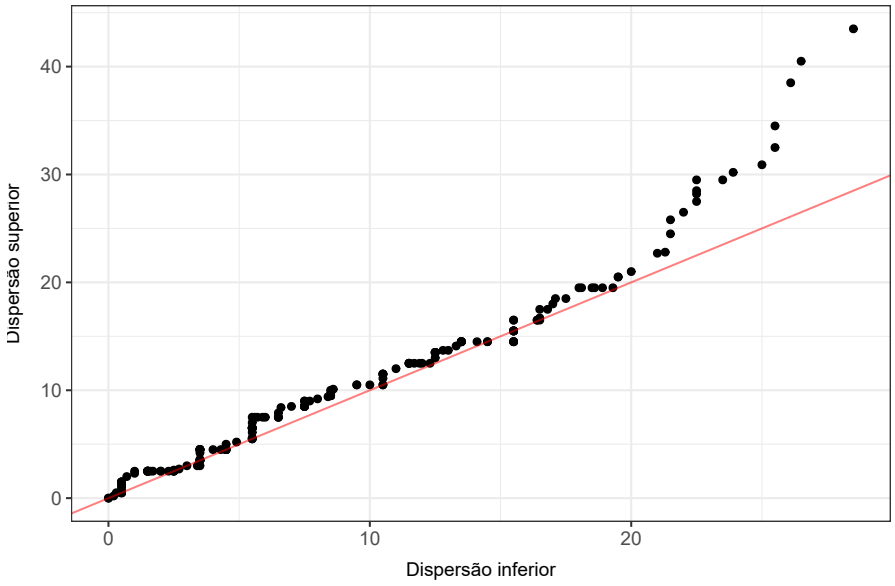
Figura 3.2: Distribuição do peso dos pacientes




```
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

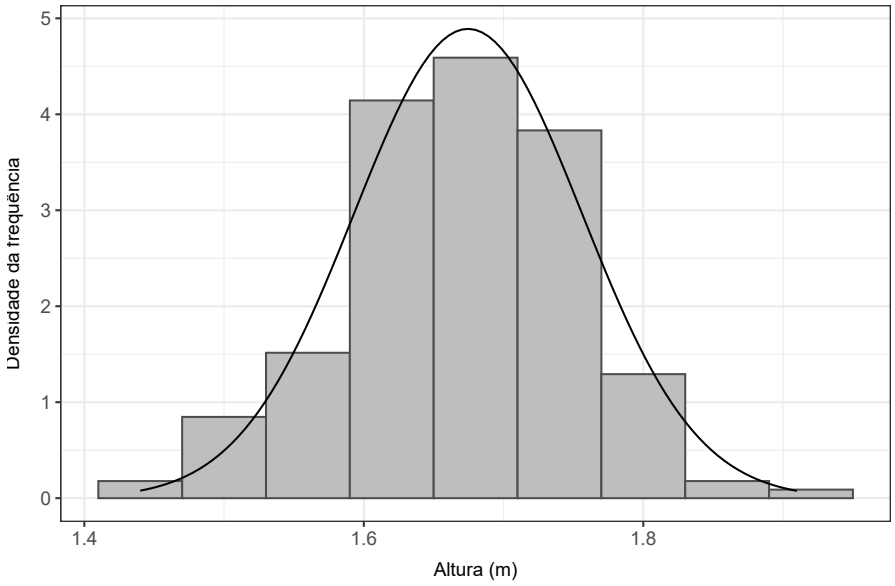



```
rehabcardio %>%
  ggplot(aes(x = Altura)) +
    geom_histogram(
      mapping = aes(y = ..density..),
      binwidth = 0.06,
      fill = "grey",
      color = "grey30"
    ) +
    geom_function(
      fun = dnorm,
      args = list(
                mean = mean(rehabcardio$Altura, na.rm = TRUE),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

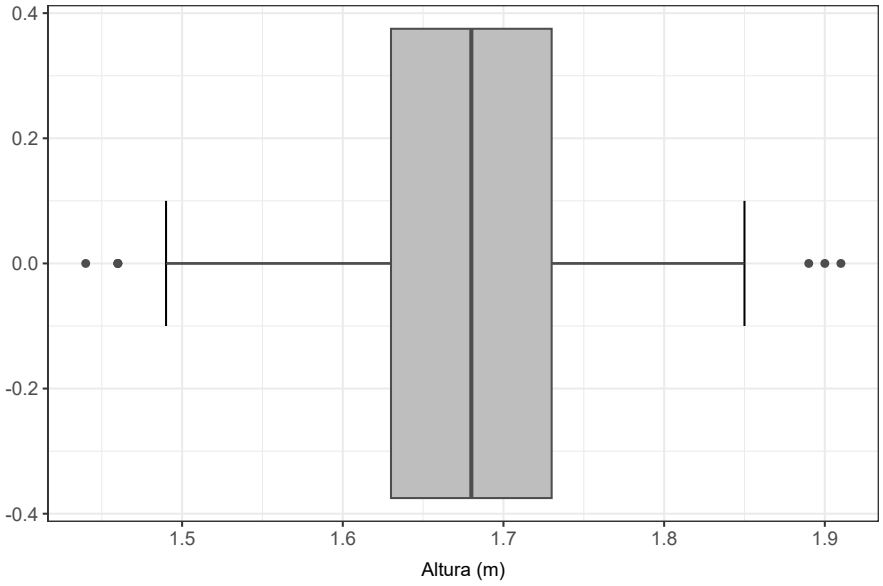
Figura 3.4: Distribuição de frequência das alturas dos pacientes




```
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
```

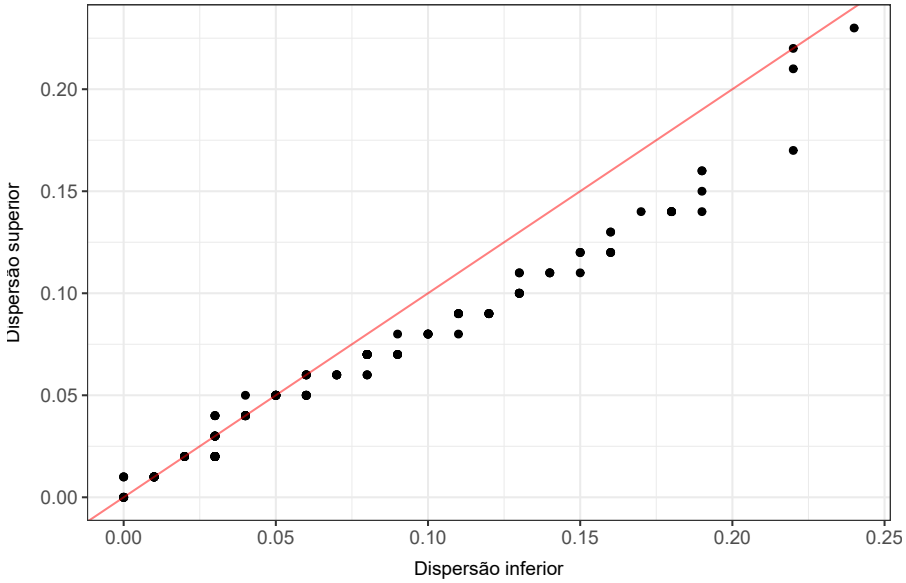
```
## Warning: Removed 7 rows containing non-finite values (`stat_boxplot()`).
## Removed 7 rows containing non-finite values (`stat_boxplot()`).
```

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
```

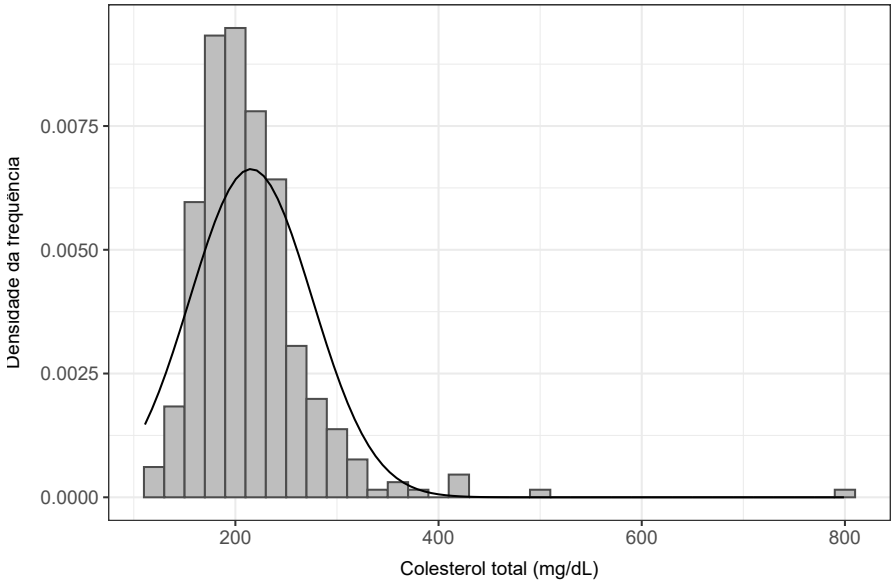
Figura 3.6: Gráfico de simetria para a altura dos pacientes




```
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()`).
```

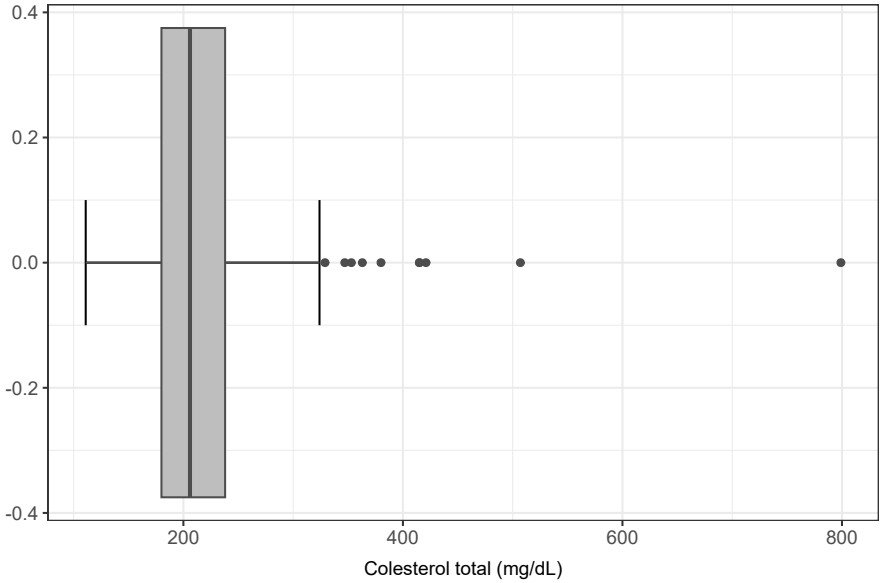
Figura 3.7: Distribuição de frequência do colesterol total dos pacientes




```
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
```

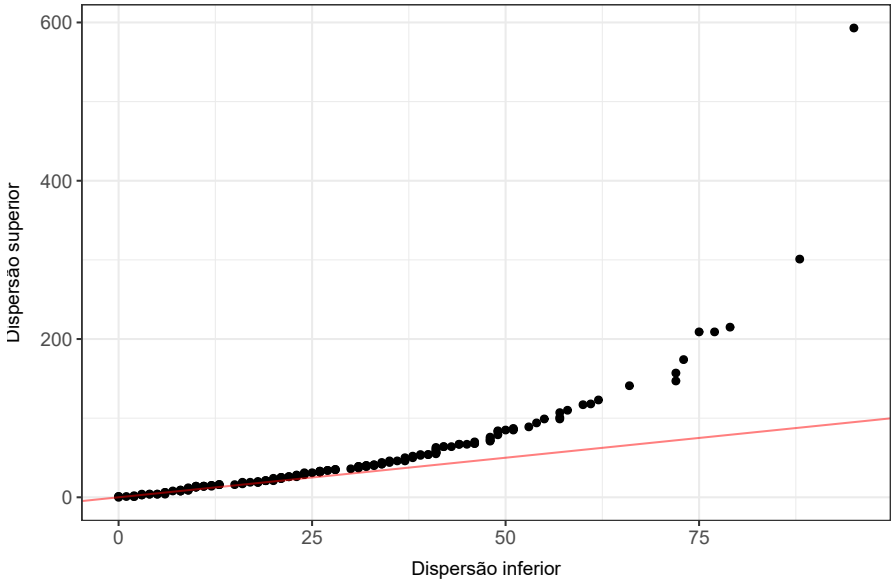
```
## Warning: Removed 54 rows containing non-finite values (`stat_boxplot()`).
## Removed 54 rows containing non-finite values (`stat_boxplot()`).
```

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
```

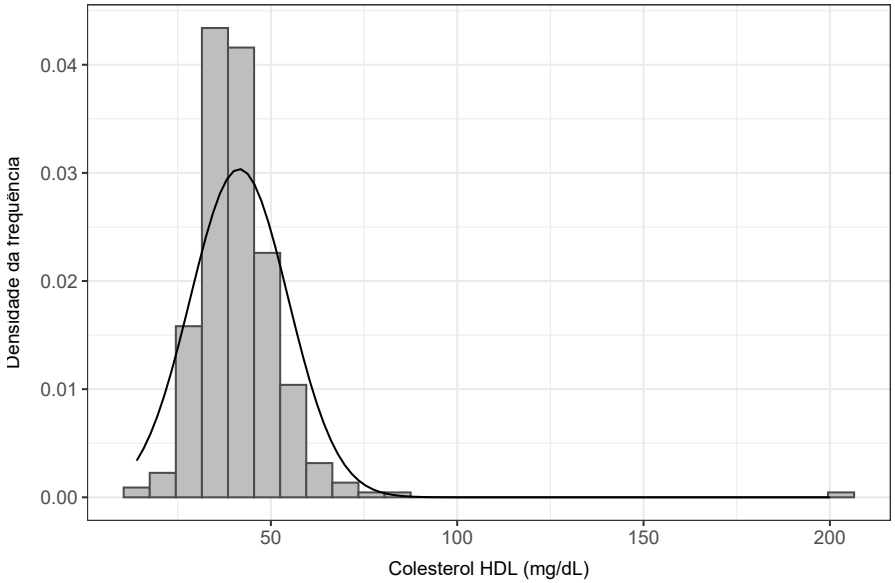
Figura 3.10: Gráfico de simetria para os níveis de colesterol total dos pacientes




```
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()`).
```

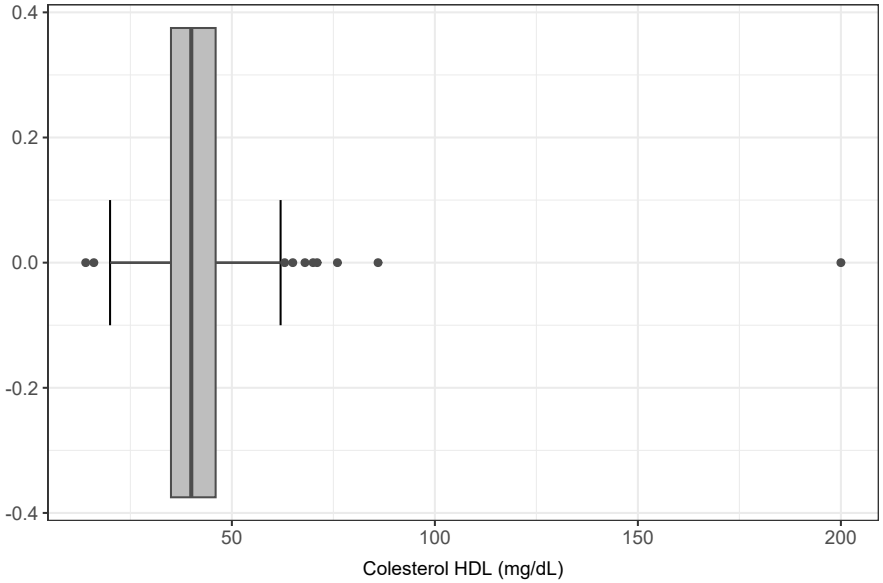
Figura 3.11: Distribuição de frequência do colesterol HDL dos pacientes




```
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
```

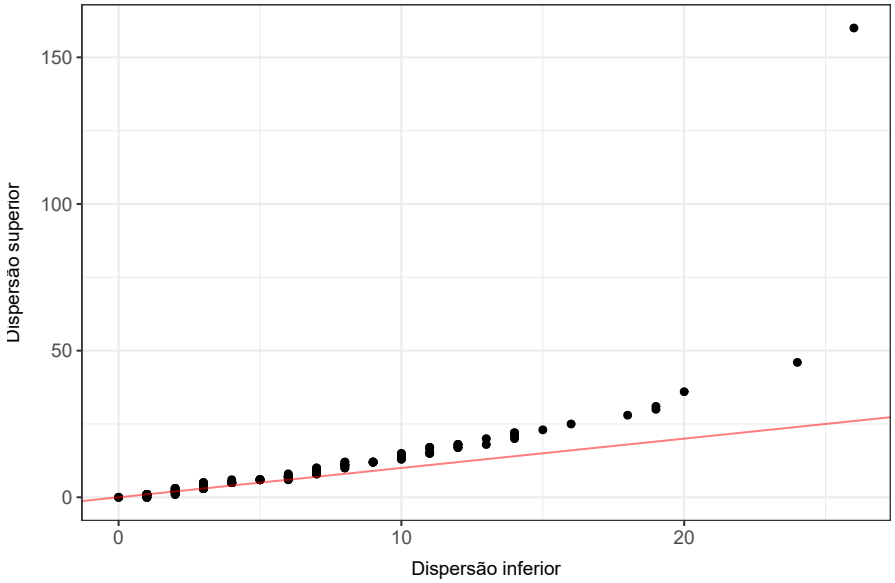
```
## Warning: Removed 65 rows containing non-finite values (`stat_boxplot()`).
## Removed 65 rows containing non-finite values (`stat_boxplot()`).
```

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
```

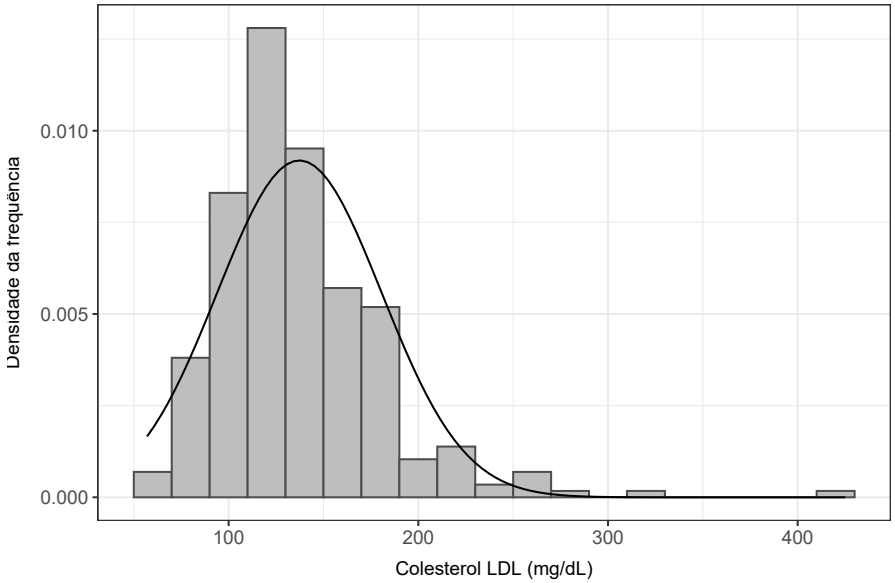
Figura 3.13: Gráfico de simetria para os níveis de colesterol HDL dos pacientes




```
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()`).
```

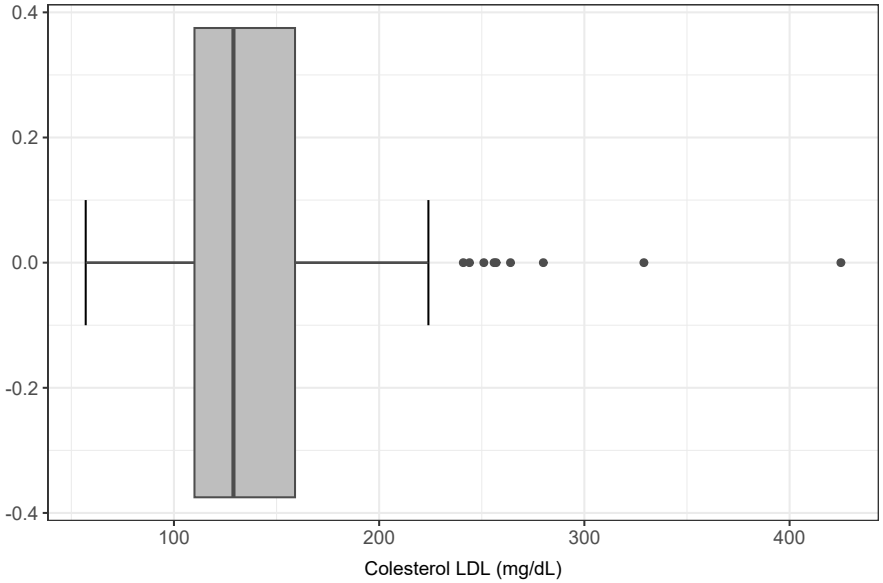
Figura 3.14: Distribuição de frequência do colesterol LDL dos pacientes




```
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
```

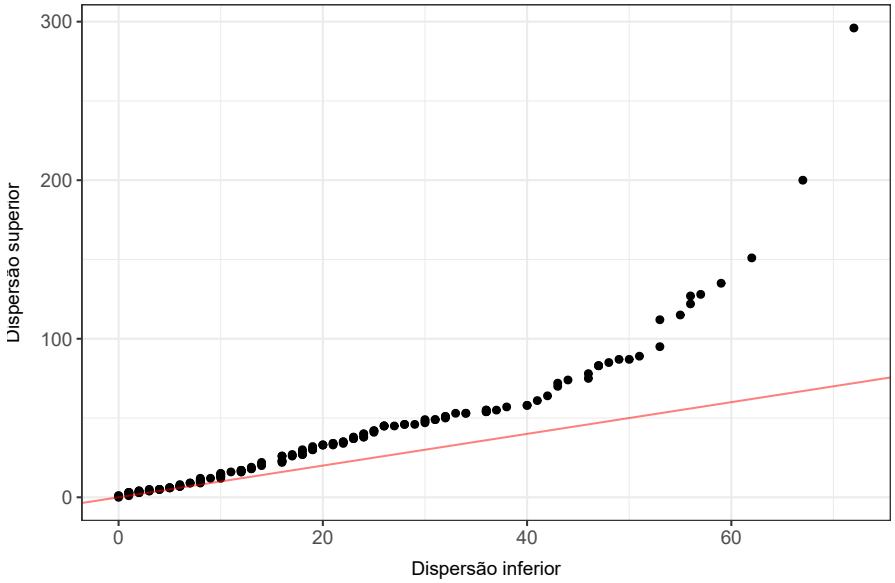
```
## Warning: Removed 92 rows containing non-finite values (`stat_boxplot()`).
## Removed 92 rows containing non-finite values (`stat_boxplot()`).
```

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 continuas

```
summ <- function(x) {
  c(
    n = sum(!is.na(x), na.rm = TRUE),
    `Mínimo` = min(x, na.rm = TRUE),
    Q1 = quantile(x, 0.25, na.rm = TRUE)[[1]],
    `Mediana` = median(x, na.rm = TRUE),
    Q3 = quantile(x, 0.75, na.rm = TRUE)[[1]],
    `Máximo` = max(x, na.rm = TRUE),
    `Média` = mean(x, na.rm = TRUE),
    `Desvio Padrão` = sd(x, na.rm = TRUE),
    `Distância Interquartil` = IQR(x, na.rm = TRUE)
  )
}

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

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

	n	Mínimo	Q1	Mediana	Q3	Máximo	Média	Desvio Padrão	Distância Interquartil
Peso	375	47,00	67,05	75,50	84,95	119,00	76,11	12,79	17,9
Altura	374	1,44	1,63	1,68	1,73	1,91	1,67	0,08	0,1
Coltot	327	111,00	180,00	206,00	238,00	799,00	215,43	60,16	58,0
HDL	316	14,00	35,00	40,00	46,00	200,00	41,52	13,14	11,0
LDL	289	57,00	110,00	129,00	159,00	425,00	137,31	43,41	49,0

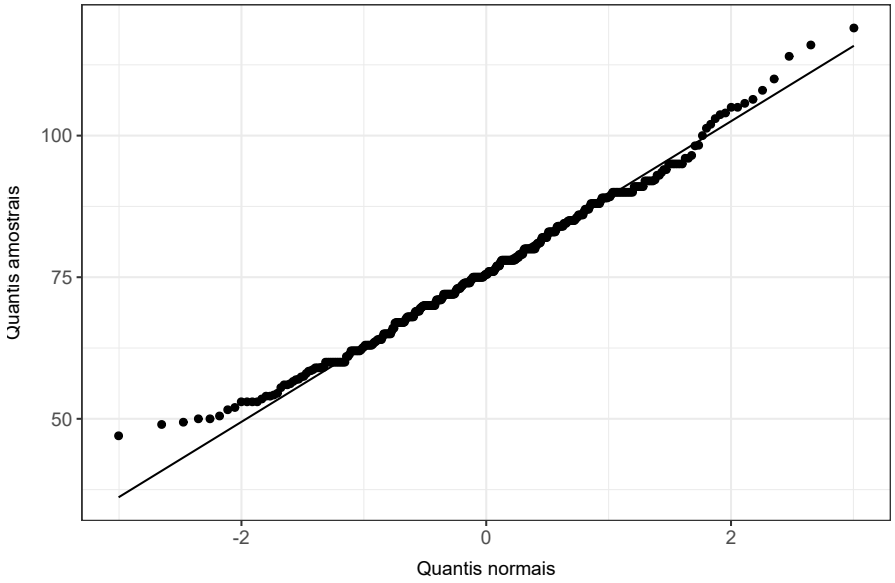
d) Avalie a compatibilidade de distribuições normais para as variáveis continuas por meio de gráficos QQ

```
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

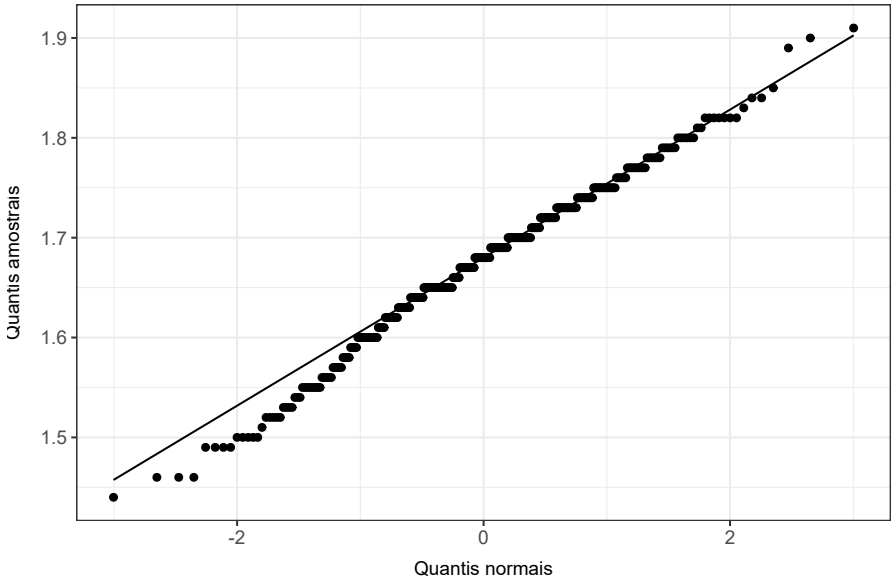



```
rehabcardio %>%
  ggplot(aes(sample = Altura)) +
    geom_qq() +
    geom_qq_line() +
    labs(
      title = "Figura 3.18: Gráfico QQ normal para a altura dos pacientes",
      x = "Quantis normais",
      y = "Quantis amostrais"
    ) +
    tema
```

```
## 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

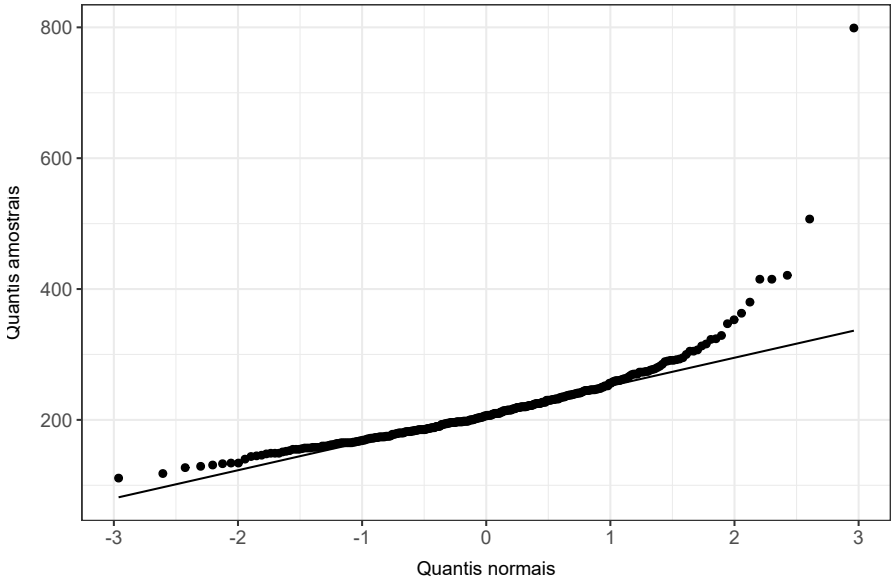



```
rehabcardio %>%
  ggplot(aes(sample = Coltot)) +
    geom_qq() +
    geom_qq_line() +
    labs(
      title = "Figura 3.19: Gráfico QQ normal para os níveis de colesterol total dos pacientes",
      x = "Quantis normais",
      y = "Quantis amostrais"
    ) +
    tema
```

```
## 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

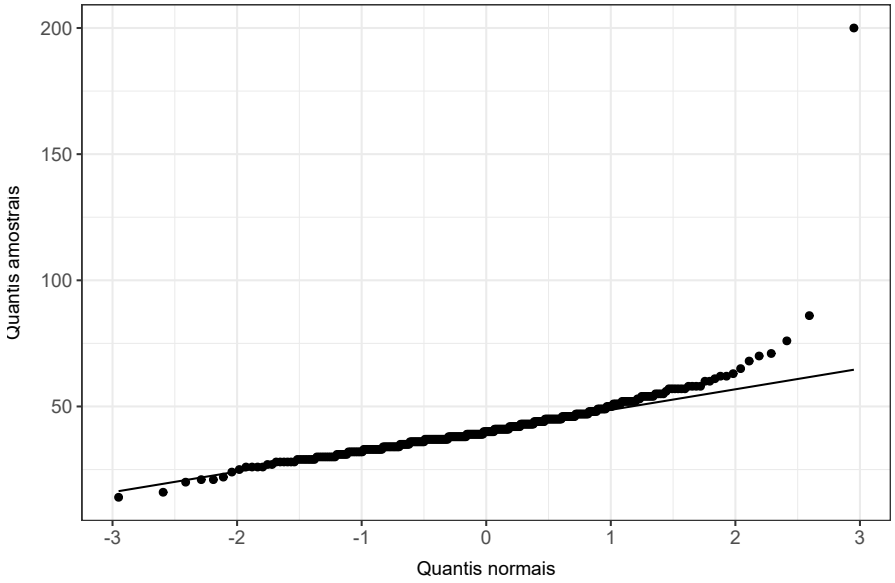



```
rehabcardio %>%
  ggplot(aes(sample = HDL)) +
    geom_qq() +
    geom_qq_line() +
    labs(
      title = "Figura 3.20: Gráfico QQ normal para os níveis de colesterol HDL dos pacientes",
      x = "Quantis normais",
      y = "Quantis amostrais"
    ) +
    tema
```

```
## 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

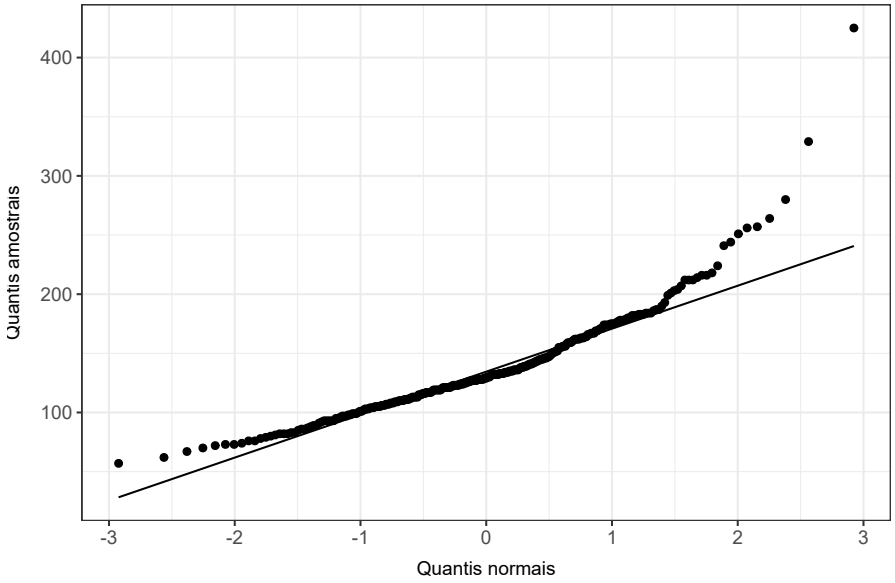



```
rehabcardio %>%
  ggplot(aes(sample = LDL)) +
    geom_qq() +
    geom_qq_line() +
    labs(
      title = "Figura 3.21: Gráfico QQ normal para os níveis de colesterol LDL dos pacientes",
      x = "Quantis normais",
      y = "Quantis amostrais"
    ) +
    tema
```

```
## 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`.

a) Construa uma tabela com as medidas de posição e dispersão estudadas para as variáveis desse arquivo.
b) 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()

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

	n	Mínimo	Q1	Mediana	Q3	Máximo	Média	Desvio Padrão	Distância Interquartil
idade	2452	19,00	58,00	69,00	81,00	110,00	68,43	15,14	23,00
tmunic	2452	0,00	39,00	50,00	62,00	103,00	49,27	20,63	23,00
htransp	2452	0,00	0,00	0,83	2,00	12,00	1,78	2,54	2,00
cargatabag	2452	0,00	0,00	1,50	41,00	256,00	25,57	39,31	41,00
antracose	2350	0,00	0,07	0,17	0,29	0,83	0,20	0,16	0,22
ses	2452	-1,00	-0,51	-0,31	-0,05	1,00	-0,24	0,36	0,47
densid	2452	0,00	0,01	0,02	0,02	0,04	0,02	0,01	0,01
distmin	2452	0,03	45,24	114,12	225,95	1964,02	170,67	190,87	180,71

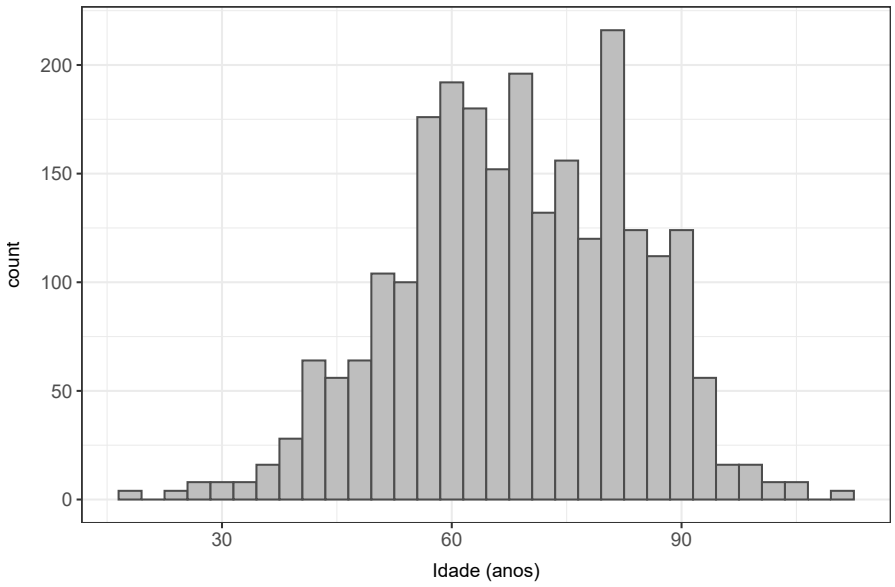
Construa histogramas e boxplots para essas variáveis e verifique que transformação é necessária para tornar mais simétricas aquelas em que a simetria pode ser questionada

Para a variável `idade`, temos o seguinte:

```
antracose %>%
  ggplot(aes(x = idade)) +
    geom_histogram(
      binwidth = 3,
      fill = "grey",
      color = "grey30"
```

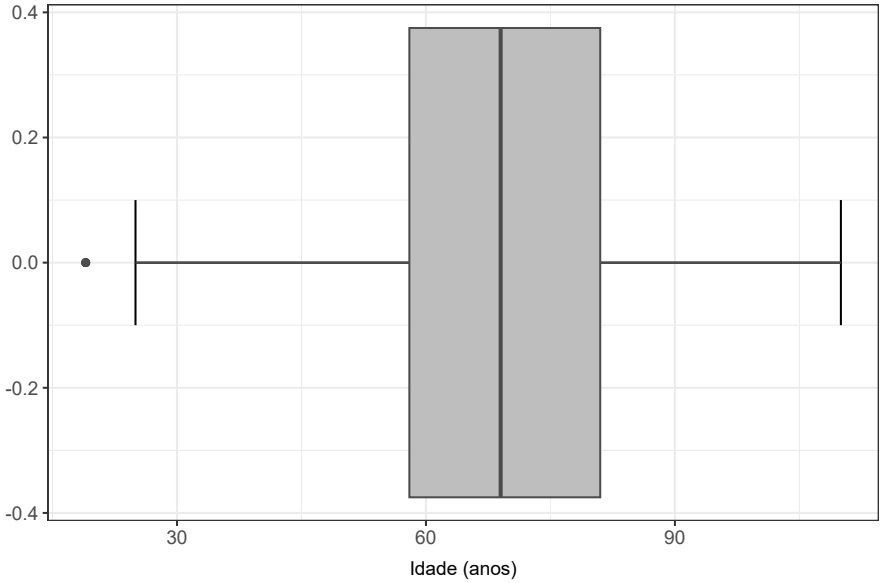
```
  ) +
  labs(
    title = "Figura 3.22: Distribuição da idade dos pacientes",
    x = "Idade (anos)"
  ) +
  tema
```

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]]
```

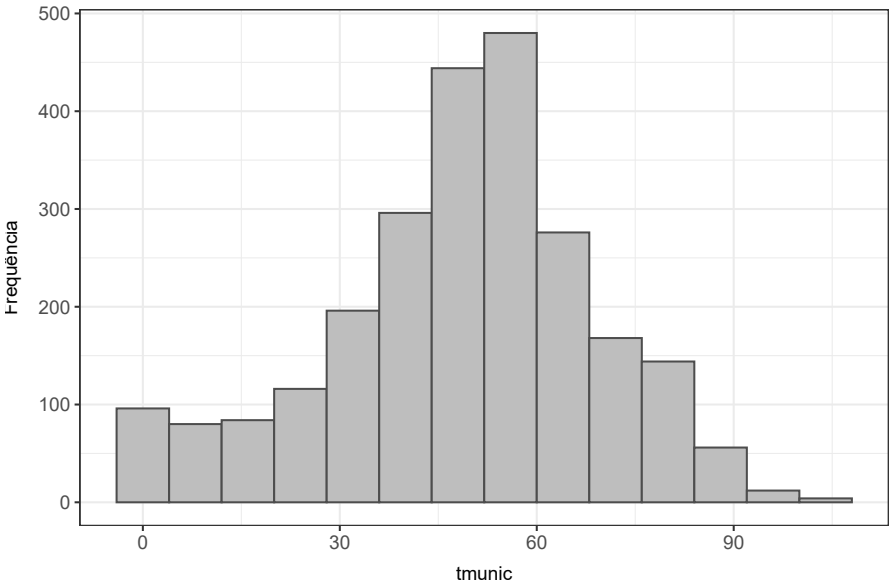
```
## [1] -1
```

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:

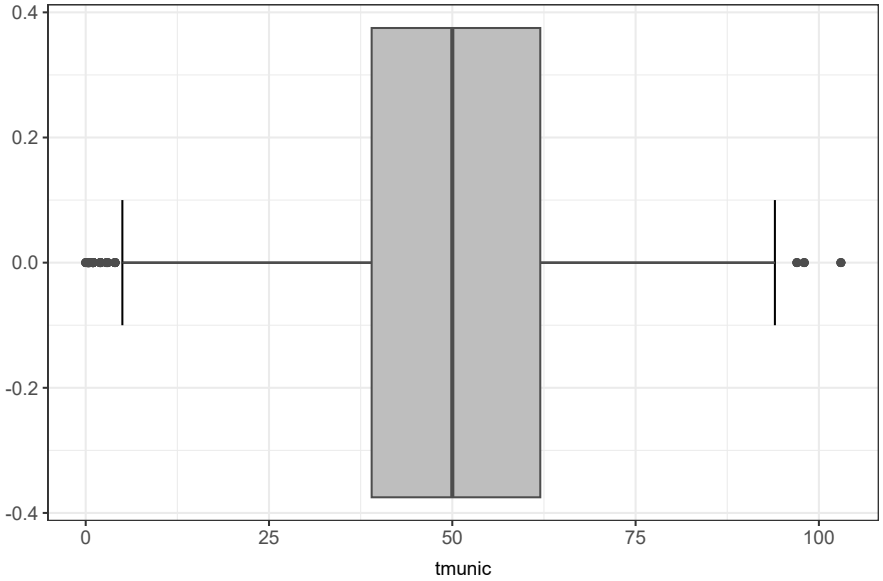
```
antracose %>%
  ggplot(aes(x = tmunic)) +
    geom_histogram(
      binwidth = 8,
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.24: Distribuição da variável `tmunic`",
      y = "Frequência"
    ) +
    tema
```

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]]
```

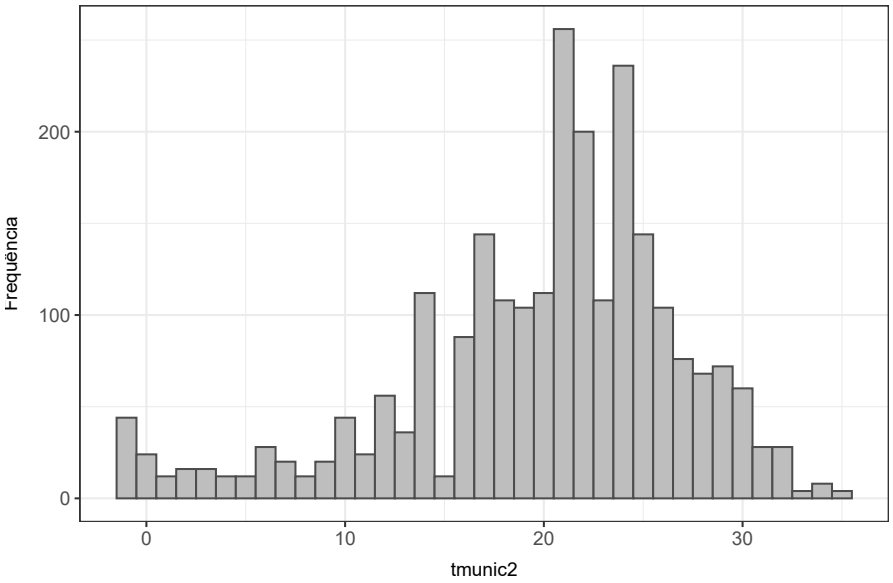
```
## [1] 0.7
```

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
```

Figura 3.26: Distribuição da variável `tmunic2`



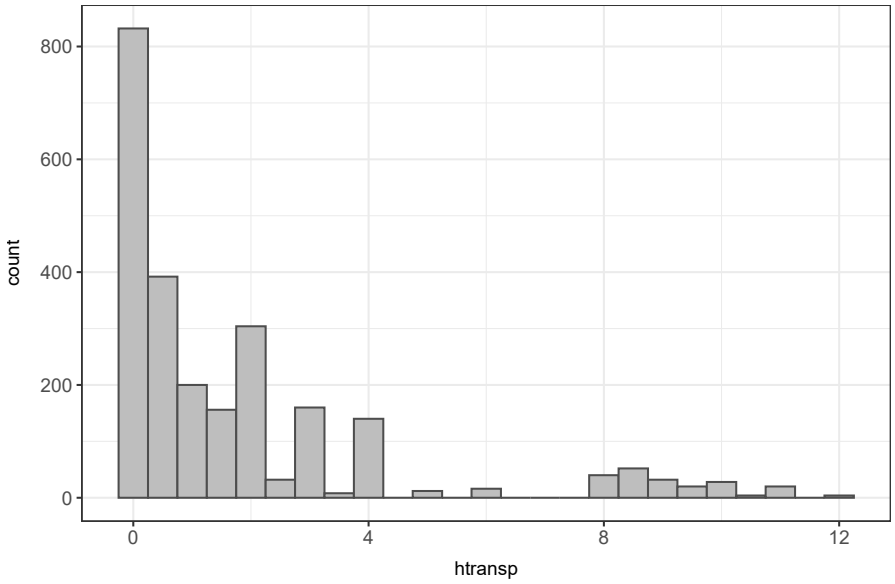
Neste caso, a transformação não melhorou a simetria dos dados. Em vez disso, o coeficiente de assimetria saltou de -0.3960939 para -0.9443456.

Para a variável `htransp`, temos:

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

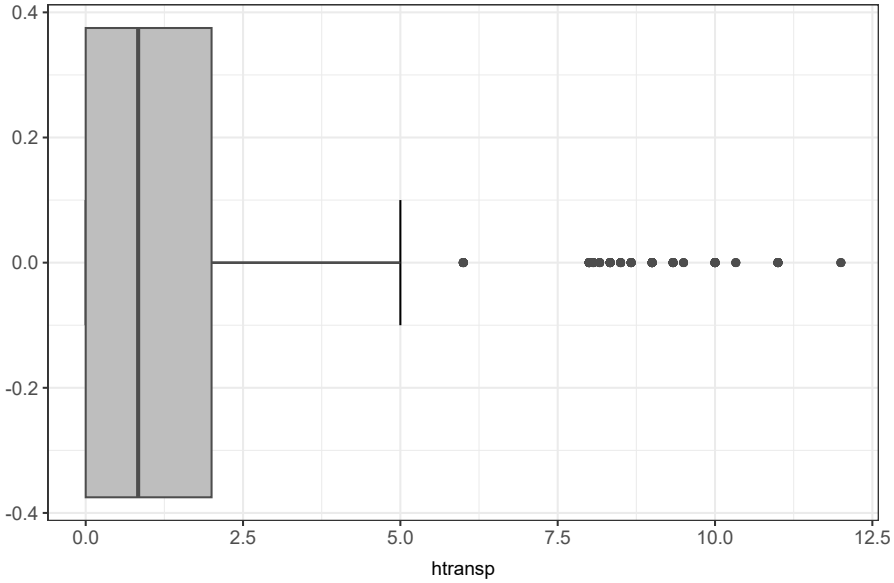
```
    labs(
      title = "Figura 3.27: Distribuição da variável `htransp`"
    ) +
    tema
```

Figura 3.27: Distribuição da variável `htransp`




```
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]]
```

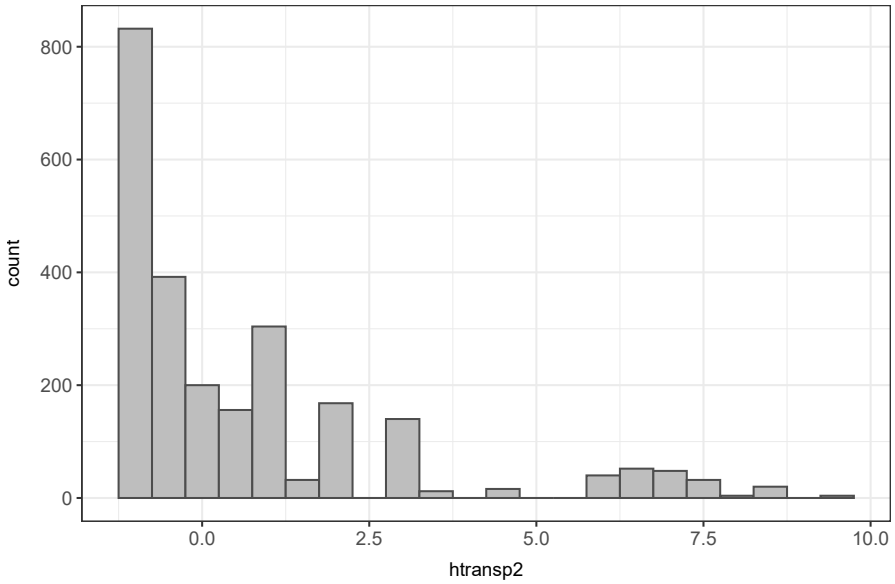
```
## [1] 0.9
```

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
```

Figura 3.29: Distribuição da variável `log(htransp)`



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
```

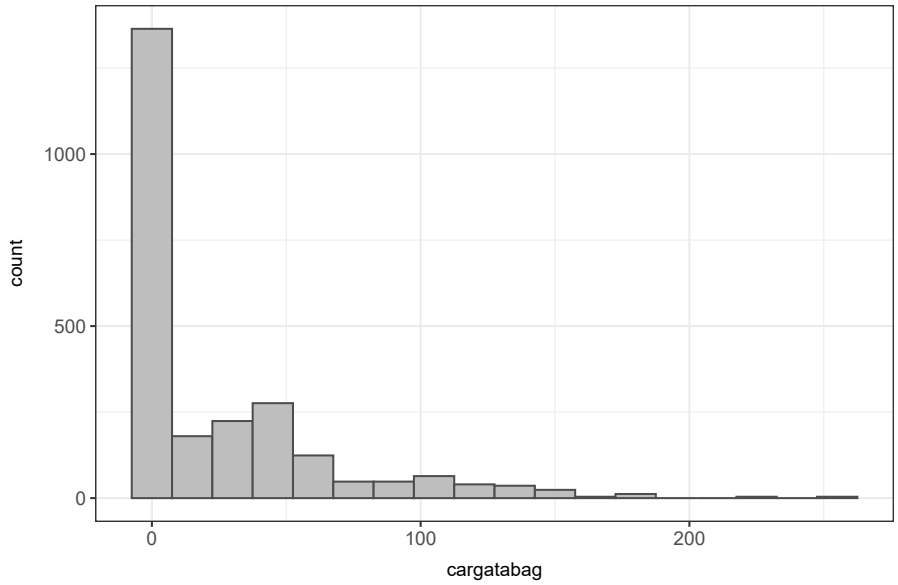
```
e1071::skewness(antracose$htransp2, type = 2)
```

```
## [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
```

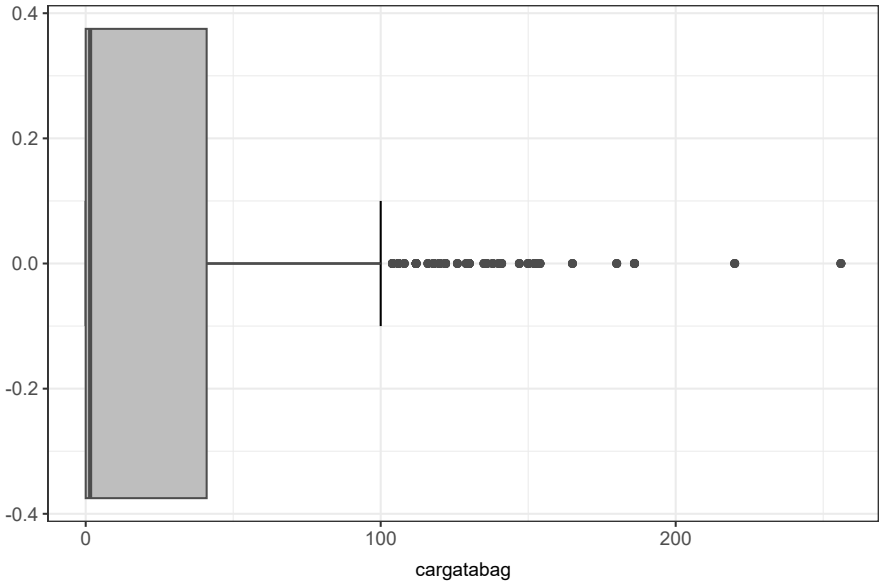
Figura 3.30: Distribuição da variável `cargatabag`




```
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
```

Figura 3.31: Distribuição da variável `cargatabag`



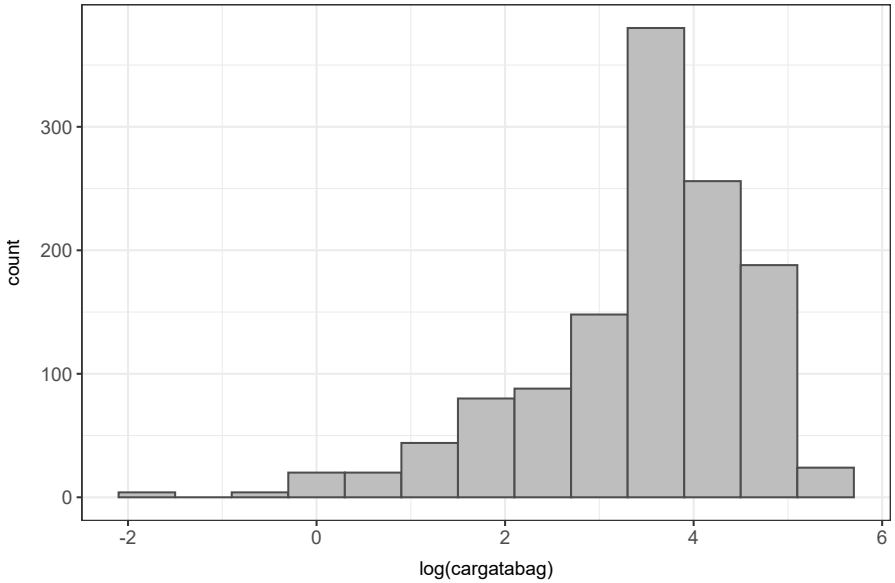
Temos uma distribuição fortemente assimétrica à direita. O Coeficiente de Assimetria de Fisher-Pearson Ajustado nos indica um grau de assimetria na ordem de 2.0334838.

Como a distribuição tem um formato exponencial, apostaremos que a transformação que leva x em $log(x)$ seja suficiente para melhorar a distibuição 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
```

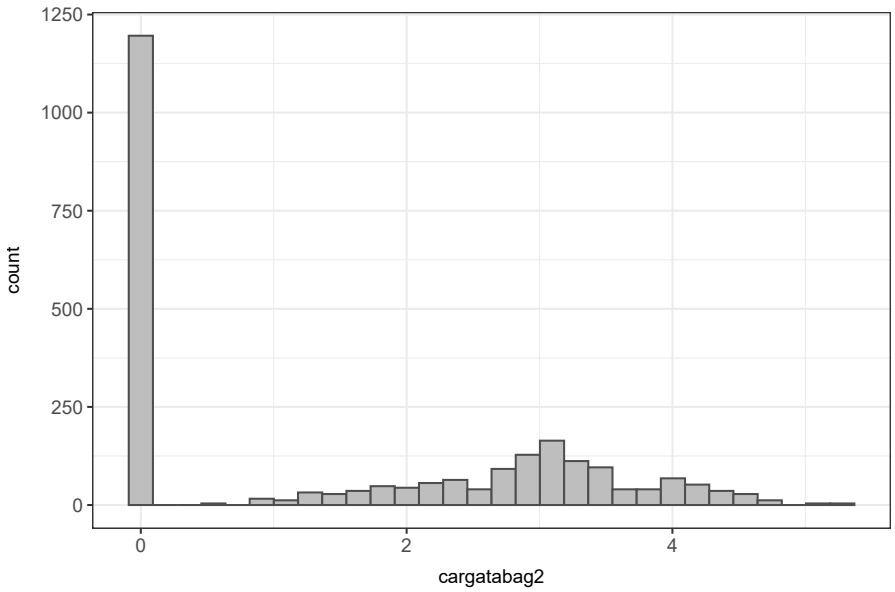
Notamos que para $p = 0.3$, nossa métrica é minimizada.

```
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`.
```

Figura 3.33: Distribuição da variável `cargatabag`



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.

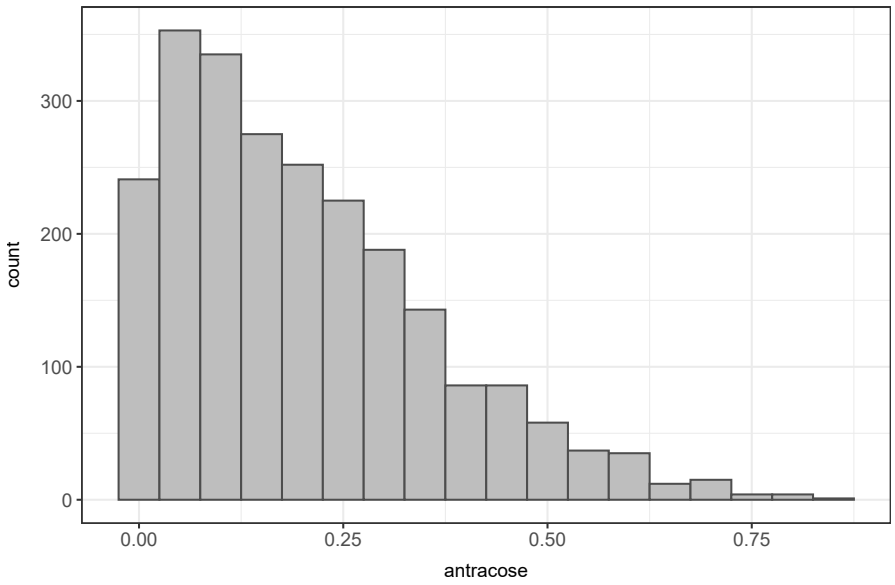
Analisemos agora a variável `antracose`:

```
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()`).
```

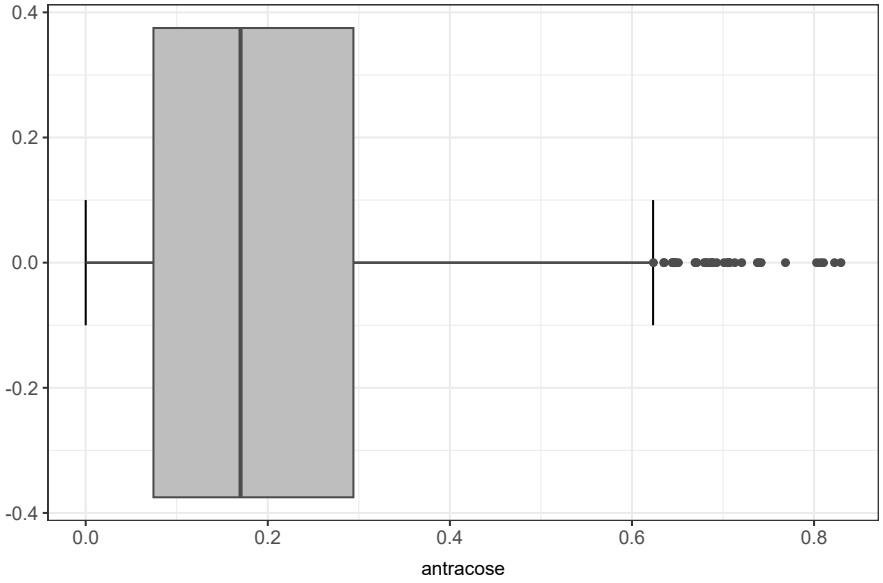
Figura 3.34: Distribuição da variável `antracose`




```
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
```

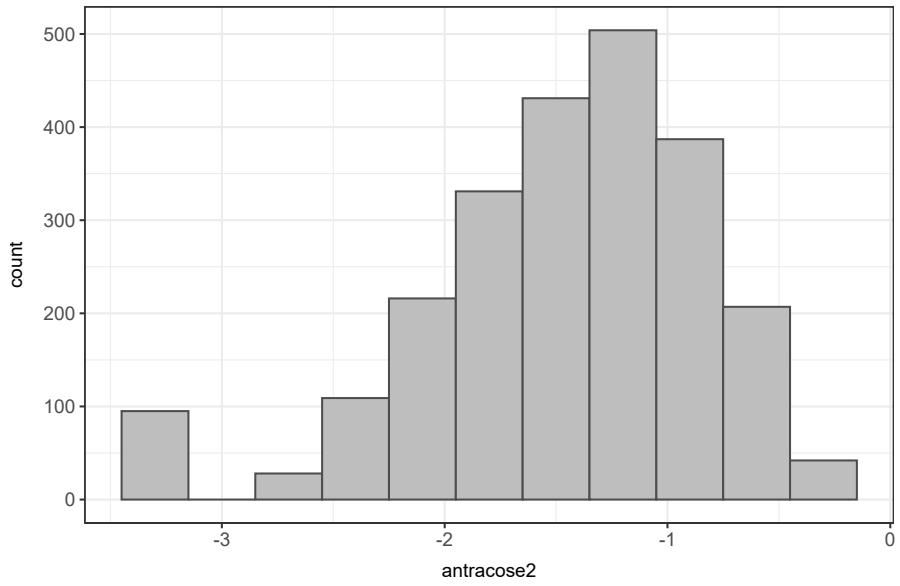
O valor $p = 0.3$ minimiza a nossa medida de avaliação. Vamos agora plotar os dados transformados e avaliar novamente o coeficiente de assimetria.

```
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()`).
```

Figura 3.36: Distribuição da variável `antracose`



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
```

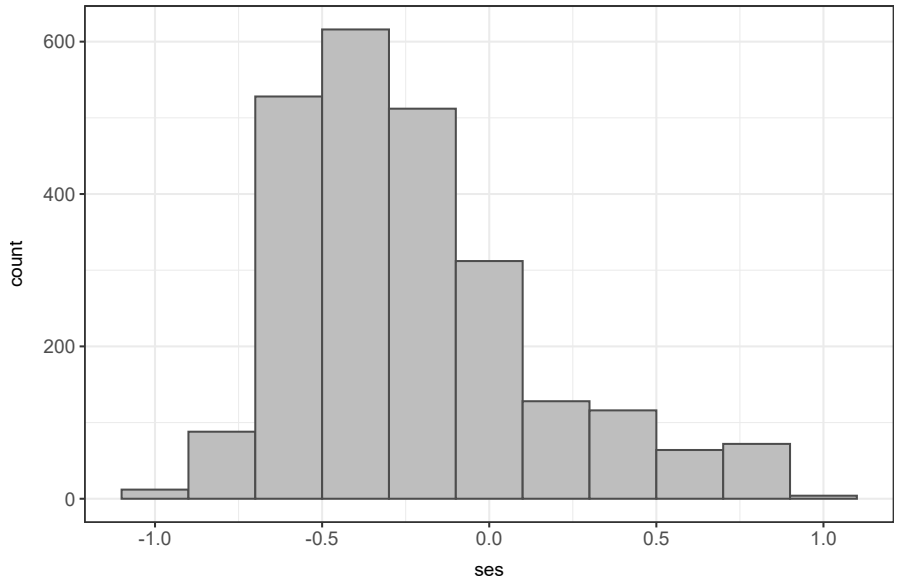
```
e1071::skewness(antracose$antracose2, na.rm = TRUE, type = 2)
```

```
## [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
```

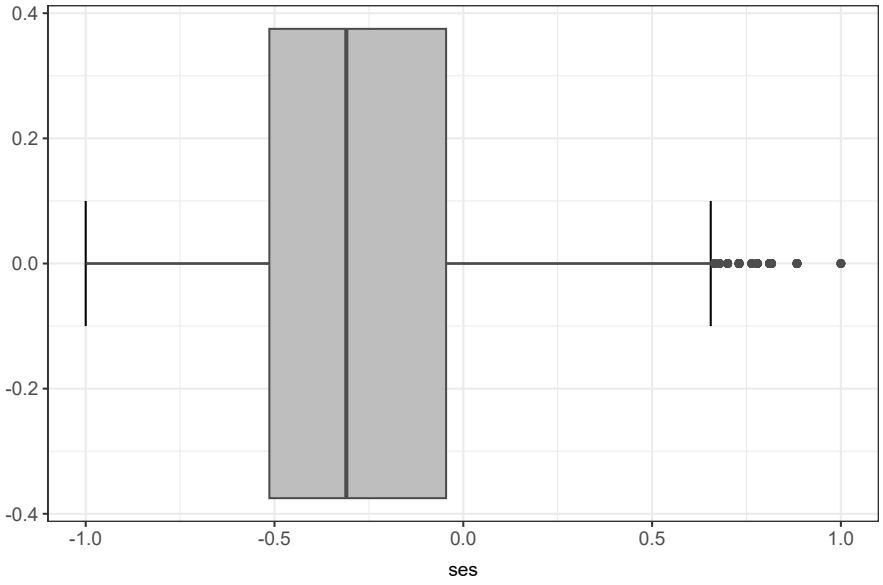
Figura 3.37: Distribuição da variável `ses`




```
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
```

Figura 3.38: Distribuição da variável `ses`



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]]
```

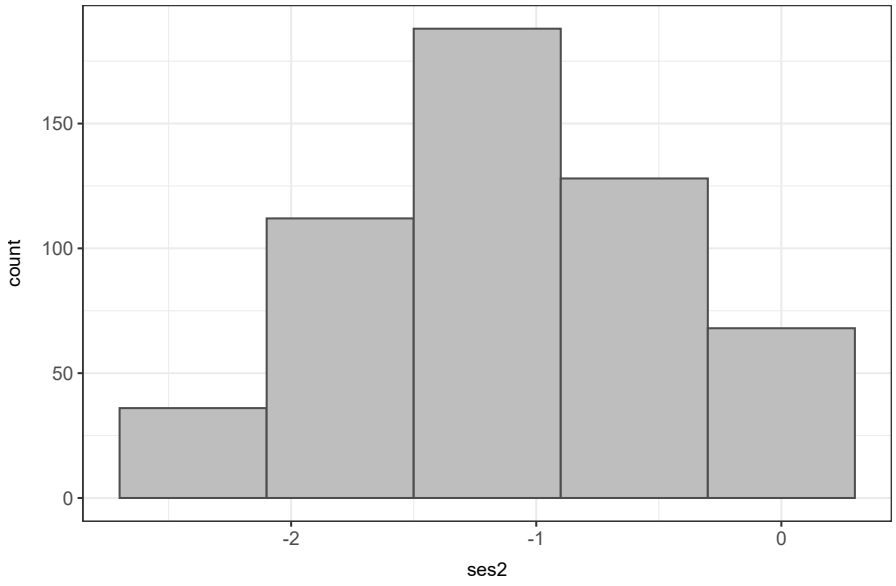
```
## [1] 0.4
```

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)
```

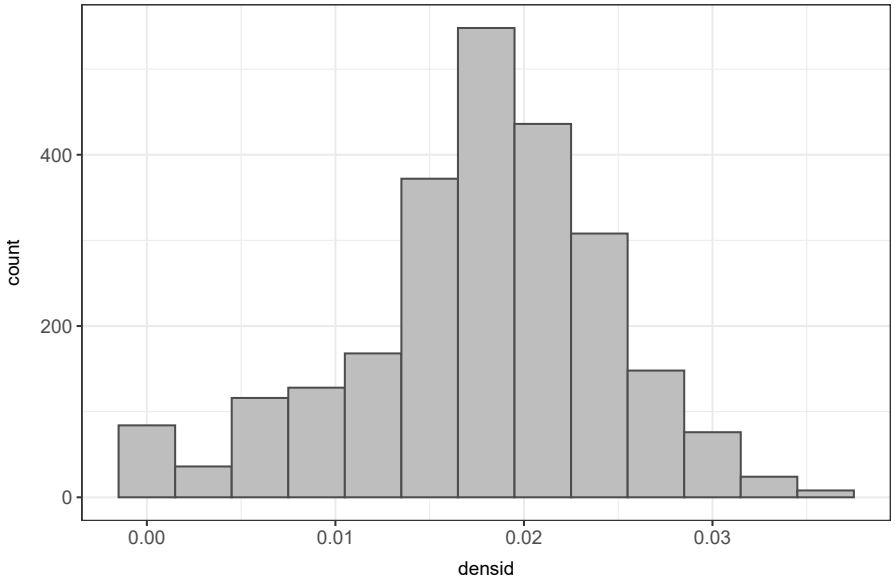
```
## [1] -0.1564781
```

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

Para a variável `densid`:

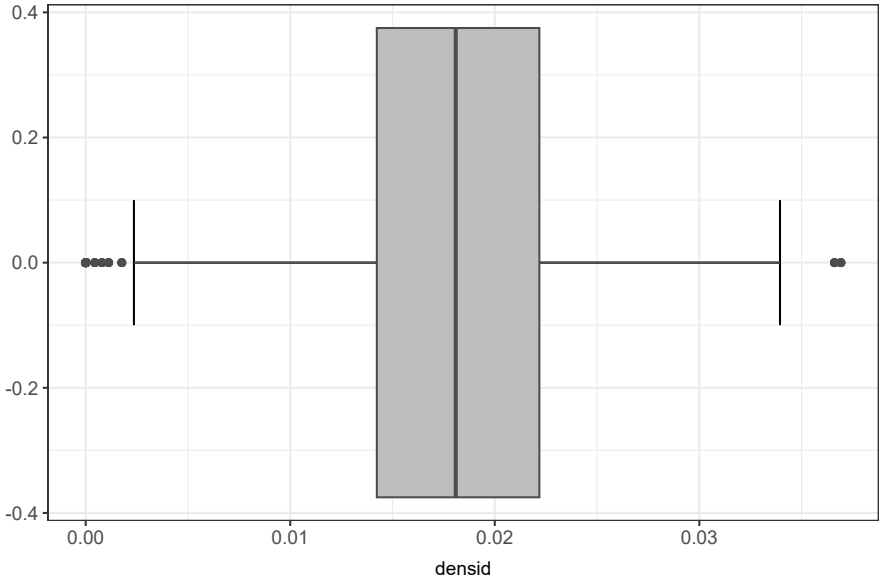
```
antracose %>%
  ggplot(aes(x = densid)) +
    geom_histogram(
      binwidth = .003,
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.40: Distribuição da variável `densid`"
    ) +
    tema
```

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
```

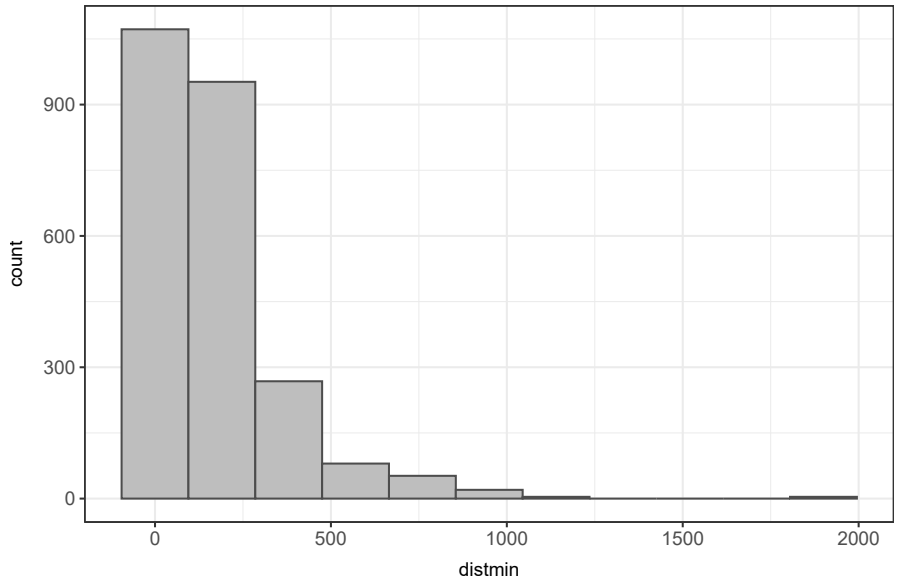
Para os diversos valores de p, o que minimiza a nossa métrica é justamente $p = 1$, que nos dá a transformação identidade. Em outras palavras, as transformações com as quais estamos trabalhando, não nos darão melhores resultados.

Por fim, vamos à variável `distmin`:

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

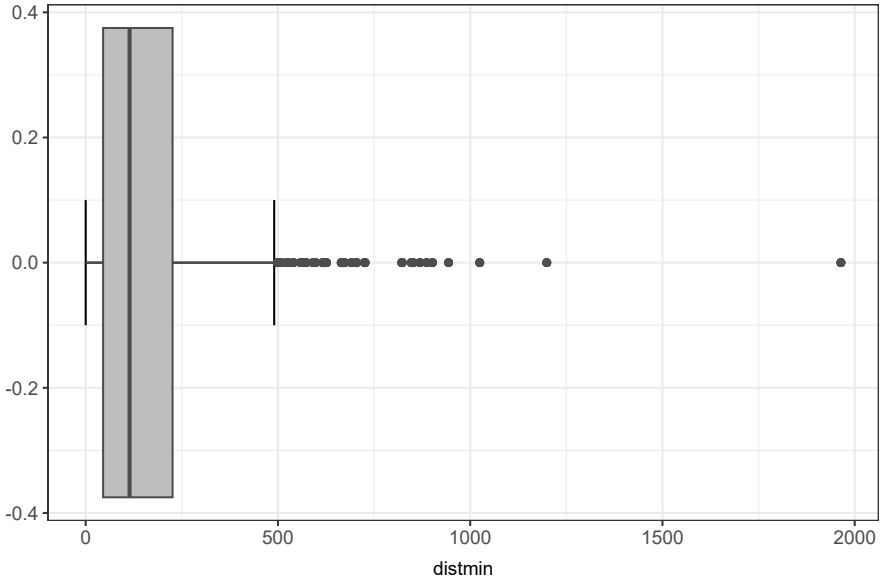
```
    geom_histogram(
      binwidth = 190,
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.42: Distribuição da variável `distmin`"
    ) +
    tema
```

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
```

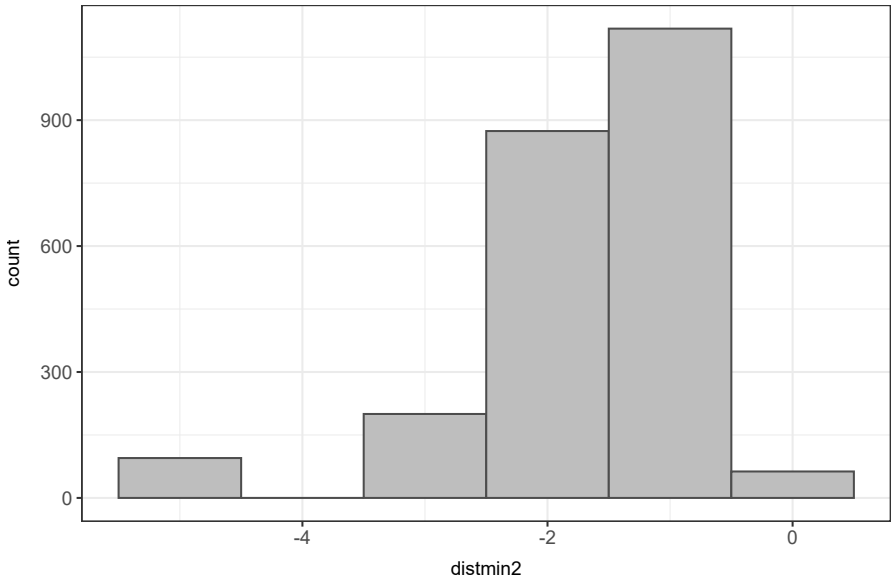
O valor $p = 0.2$ minimiza a nossa medida de avaliação. Vamos agora plotar os dados transformados e avaliar novamente o coeficiente de assimetria.

```
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()`).
```

Figura 3.44: Distribuição da variável `antracose`



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.

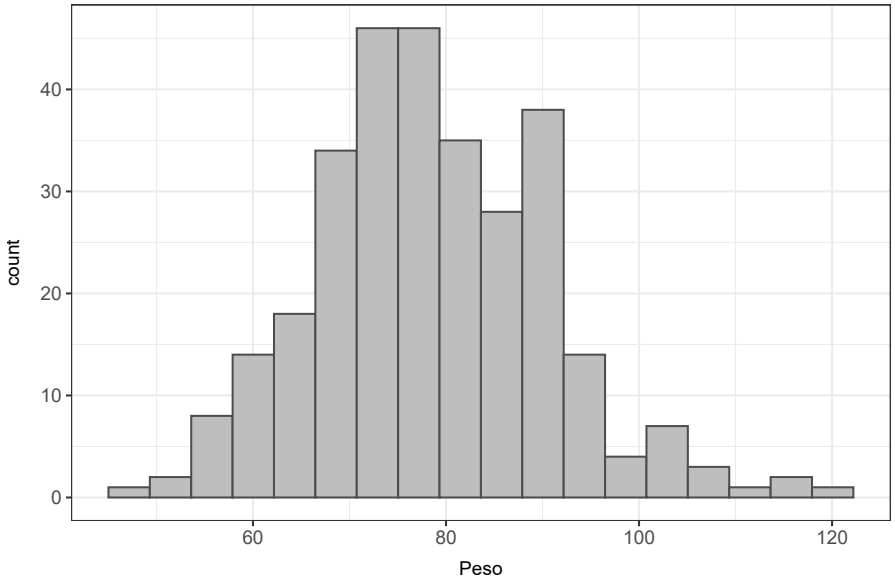
```
freedman_diaconis <- function(x) {
  s <- sd(x, na.rm = TRUE)
  n <- length(x)

  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

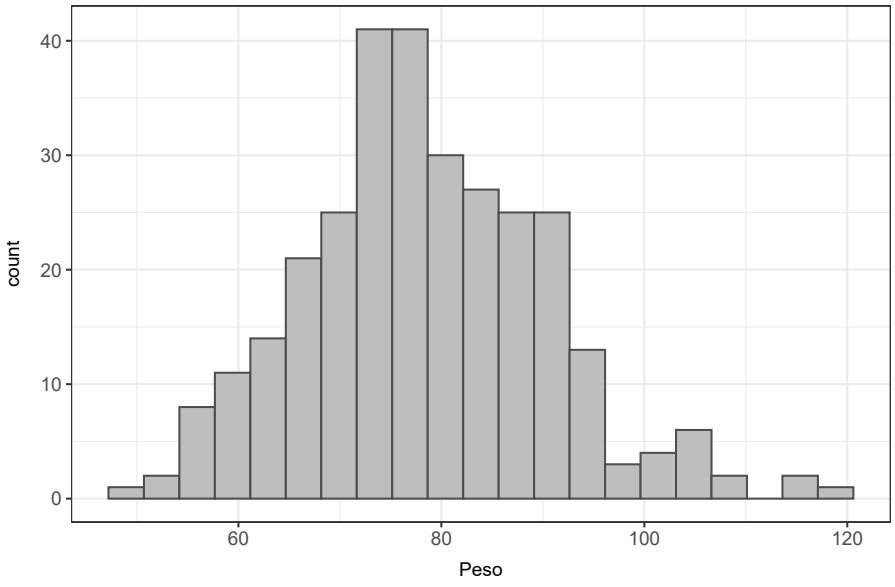



```
rehabcardio_h %>%
  ggplot(aes(Peso)) +
    geom_histogram(
      binwidth = sturges(rehabcardio_h$Peso),
      fill = "grey",
      color = "grey30"
    ) +
```

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

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

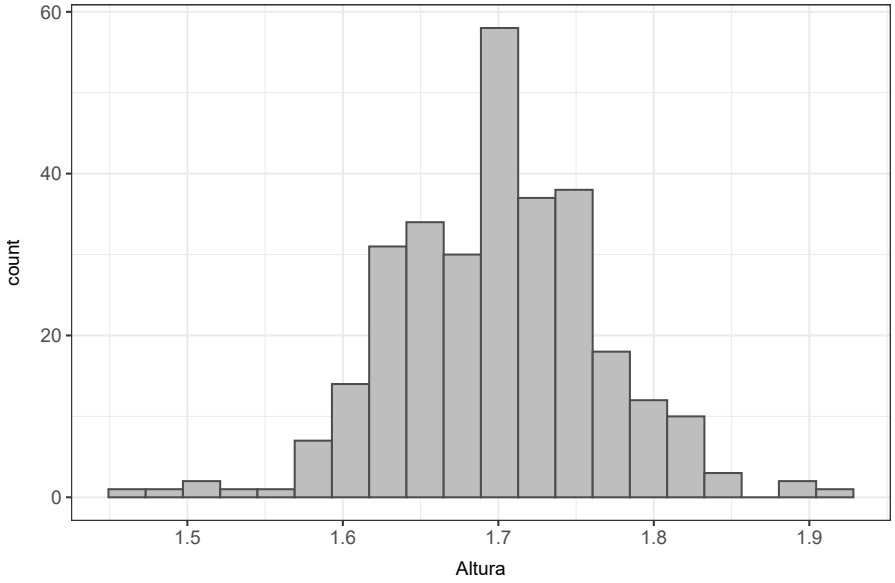
Figura 3.46: Distribuição do peso utilizando Sturges




```
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()`).
```

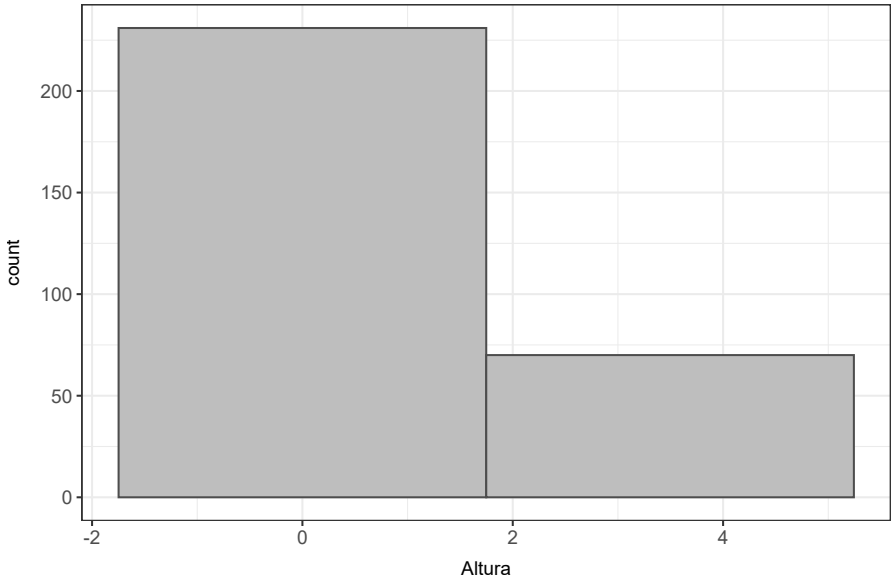
Figura 3.47: Distribuição da altura utilizando Freedman-Diaconis




```
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 atipico $61,1$, 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 atipico indicado nessas medidas.

Solução. Em primeiro lugar, carregamos o conjunto de dados:

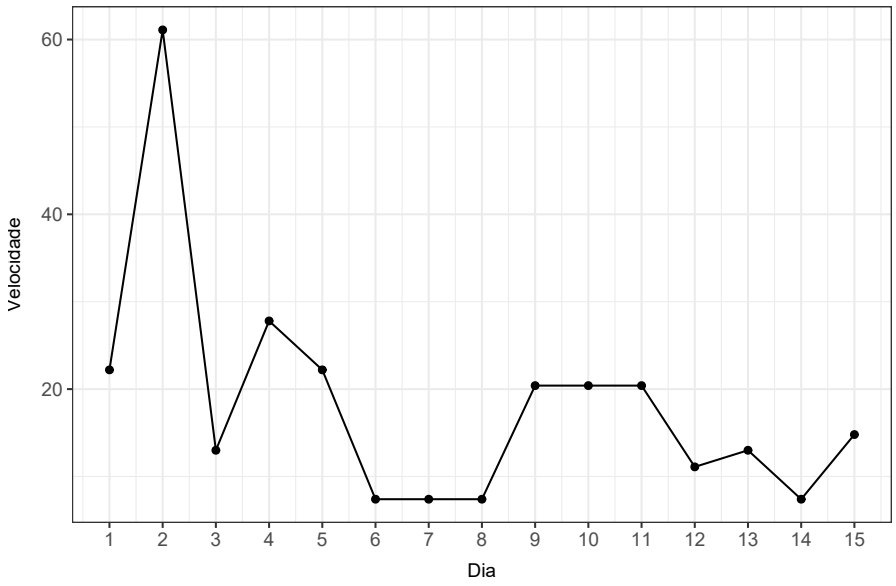
```
vento <- readxl::read_xls(paste0(data_dir, "vento.xls"))
```

Esboçaremos um gráfico mostrando a velocidade do vento a cada dia.

```
vento %>%
  ggplot(aes(t, vt)) +
    geom_line() +
    geom_point() +
    labs(
      title = "Figura 3.49: Velocidade do vento nos primeiros quinze dias de dezembro",
      x = "Dia",
      y = "Velocidade"
```

```
  ) +
  scale_x_continuous(breaks = seq(1, 15, 1)) +
  tema
```

Figura 3.49: Velocidade do vento nos primeiros quinze dias de dezembro



Medidas de posição

A média da amostra é 18.4 e o desvio padrão é 13.525373. Os quartis da amostras são apresentados abaixo:

```
t1 <- data.frame(
            `Mínimo` = min(vento$vt),
            `Q1` = quantile(vento$vt, 0.25)[[1]],
            `Mediana` = median(vento$vt),
            `Q3` = quantile(vento$vt, 0.75)[[1]],
            `Máximo` = max(vento$vt),
            `Média` = mean(vento$vt),
            `Desvio Padrão` = sd(vento$vt)
          )

vento2 <- vento %>% filter(vt < 60)

t2 <- data.frame(
```

```
            `Mínimo` = min(vento2$vt),
            `Q1` = quantile(vento2$vt, 0.25)[[1]],
            `Mediana` = median(vento2$vt),
            `Q3` = quantile(vento2$vt, 0.75)[[1]],
            `Máximo` = max(vento2$vt),
            `Média` = mean(vento2$vt),
            `Desvio Padrão` = sd(vento2$vt)
          )

resumo <- bind_rows(t1, t2, t1 - t2)
row.names(resumo) <- c("Antes", "Depois", "Diferença")

resumo %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.6:** Medidas de resumo para a velocidade do vento",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.6: Tabela 3.6: Medidas de resumo para a velocidade do vento

	Mínimo	Q1	Mediana	Q3	Máximo	Média	Desvio.Padrão
Antes	7,4	9,25	14,8	21,3	61,1	18,40	13,53
Depois	7,4	8,33	13,9	20,4	27,8	15,35	6,84
Diferença	0,0	0,92	0,9	0,9	33,3	3,05	6,69

Notamos que após remoção do valor atípico, há uma diferença considerável tanto na média, como no desvio padrão. Este caiu pela metade.

Exercício 3.5

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

Solução. Para a construção do diagrama de ramos e folhas, utilizamos a função `stem()`.

```
stem(vento$vt, scale = 2, atom = 0.001)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 7777
##   1 | 1335
##   2 | 000228
##   3 |
##   4 |
##   5 |
##   6 | 1
```

Exercício 3.6

Transforme os dados do Exercicio 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. Vamos utilizar as funções `boxcox()` e `calcular_metrica()` criadas anteriormente.

```
p <- c(0,1/4, 1/3, 1/2, 3/4)
dps <- vector("double", length(p))
for (i in seq_along(p)) {
  dps[i] <- vento$vt %>%
              boxcox(p[i]) %>%
              calcular_metrica()
}

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

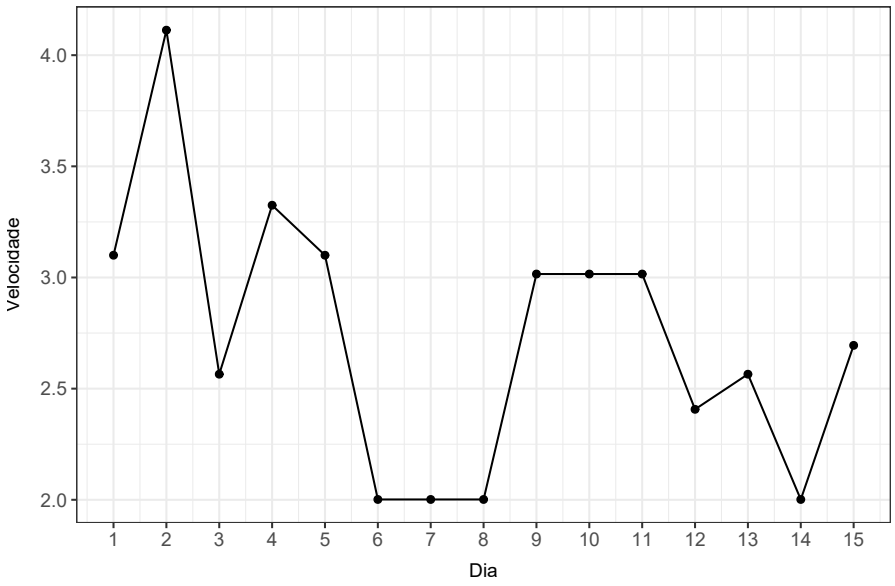
```
## [1] 0
```

A melhor opção para a transformação é $p = 0$, logo a transformação é dada pela função `log`.

```
vento %>%
  ggplot(aes(t, boxcox(vt, 0))) +
    geom_line() +
    geom_point() +
    labs(
      title = "Figura 3.50: Velocidade do vento nos primeiros quinze dias de dezembro",
      x = "Dia",
```

```
    y = "Velocidade"
  ) +
  scale_x_continuous(breaks = seq(1, 15, 1)) +
  tema
```

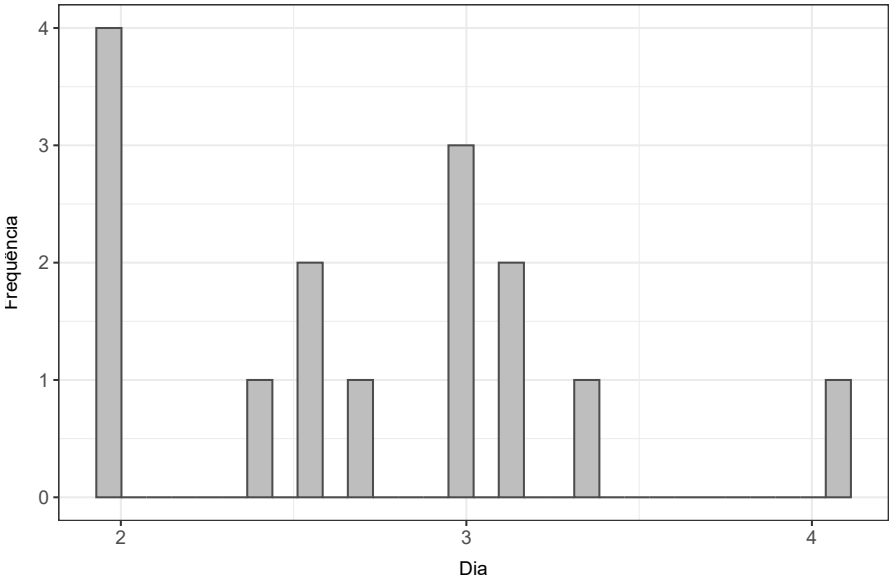
Figura 3.50: Velocidade do vento nos primeiros quinze dias de dezembro




```
vento %>%
  ggplot(aes(log(vt))) +
    geom_histogram(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.51: Distribuição da velocidade do vento",
      x = "Dia",
      y = "Frequência"
    ) +
    scale_x_continuous(breaks = seq(1, 15, 1)) +
    tema
```

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

Figura 3.51: Distribuição da velocidade do vento



Exercício 3.7

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

Solução. Inicialmente carregaos os dados:

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

E vamos verificar um resumo das variáveis:

```
summary(poluicao)
```

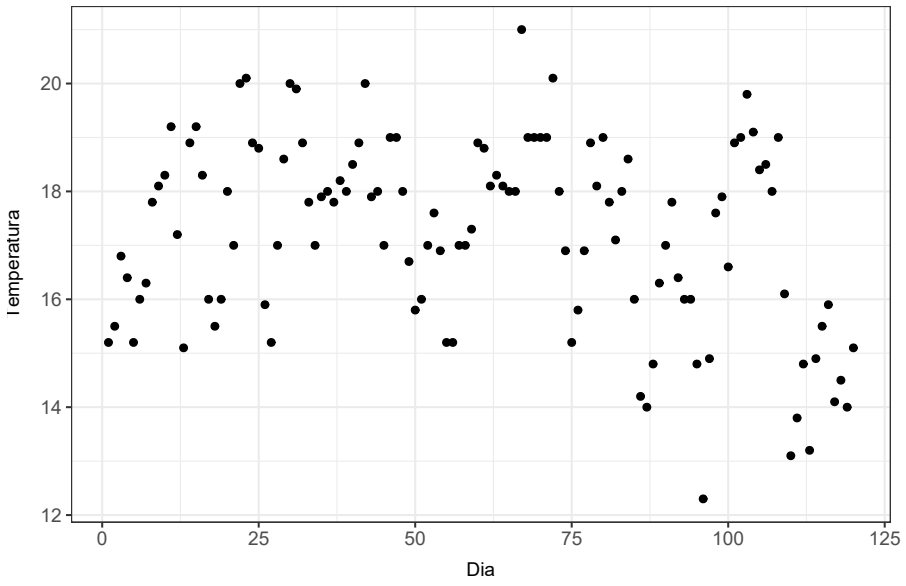
```
##      data                dia              CO              O3        
##  Length:120         Min.   :  1.00   Min.   : 4.700   Min.   :  2.70  
##  Class :character   1st Qu.: 30.75   1st Qu.: 6.300   1st Qu.: 34.40  
##  Mode  :character   Median : 60.50   Median : 7.200   Median : 62.95  
##                     Mean   : 60.50   Mean   : 7.464   Mean   : 76.33  
##                     3rd Qu.: 90.25   3rd Qu.: 8.025   3rd Qu.:114.00  
##                     Max.   :120.00   Max.   :12.500   Max.   :233.20  
##       temp            umid      
##  Min.   :12.30   Min.   :50.00  
```

```
##  1st Qu.:16.00   1st Qu.:62.00
##  Median :17.70   Median :67.50
##  Mean   :17.22   Mean   :70.78
##  3rd Qu.:18.60   3rd Qu.:78.00
##  Max.   :21.00   Max.   :99.00
```

Temos 120 observações, sendo que a temperatura é dada pela variável `temp` e tem média 17.2158333 e desvio padrão 1.7597267.

```
poluicao %>%
  ggplot(aes(dia, temp)) +
    geom_point() +
    labs(
      title = "Figura 3.52: Variação da temperatura nos primeiros 120 dias do ano",
      x = "Dia",
      y = "Temperatura"
    ) +
    tema
```

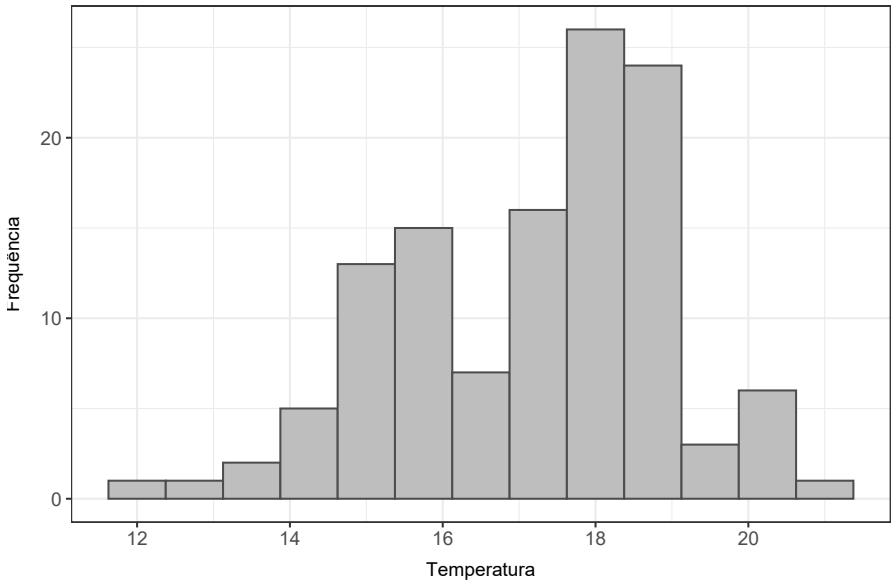
Figura 3.52: Variação da temperatura nos primeiros 120 dias do ano




```
poluicao %>%
  ggplot(aes(temp)) +
    geom_histogram(
```

```
    binwidth = 0.75,
    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = "Figura 3.53: Distribuição da temperatura",
    x = "Temperatura",
    y = "Frequência"
  ) +
  tema
```

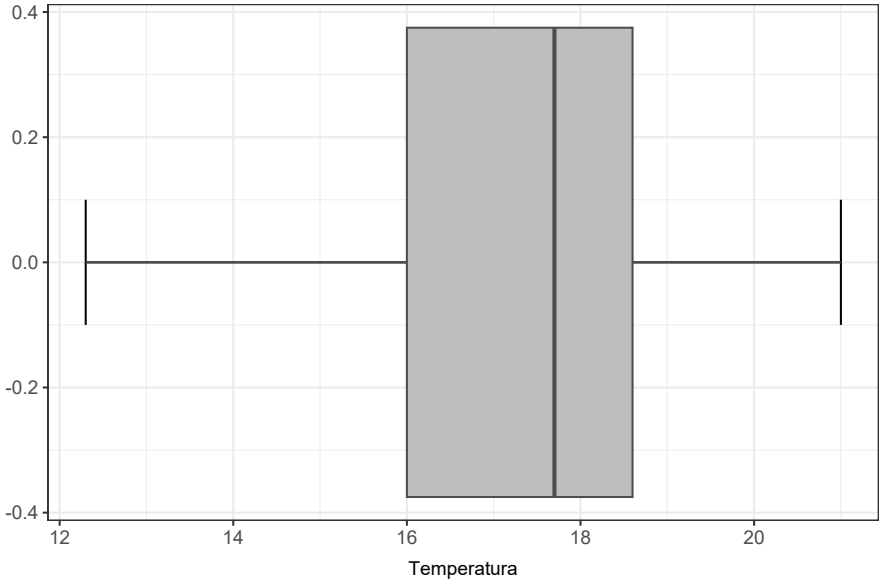
Figura 3.53: Distribuição da temperatura




```
poluicao %>%
  ggplot(aes(temp)) +
    stat_boxplot(geom = "errorbar", width = 0.2) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.54: Distribuição da temperatura",
      x = "Temperatura"
```

```
    ) +
    tema
```

Figura 3.54: Distribuição da temperatura



A distribuição tem uma leve assimetria à esquerda. O coeficiente de Assimetria de Fisher-Pearson Ajustado é -0.4253503

Utilizaremos a transformação de Box-Cox para tentar melhorar a simetria dos dados.

```
p <- seq(-3, 3, 1/10)

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

poluicao$temp2 <- boxcox(poluicao$temp, p[min_rank(dps)[1]])
```

Comparando os coeficientes de assimetria antes e após a transformação:

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

```
## [1] -0.4253503
```

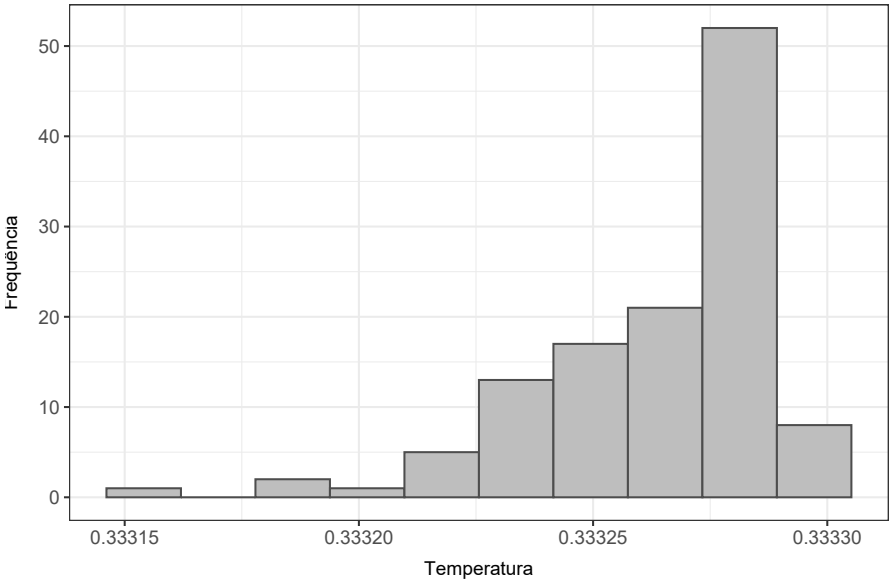
```
e1071::skewness(poluicao$temp2, na.rm = TRUE, type = 2)
```

```
## [1] -1.53657
```

Notamos que a transformação não melhorou a assimetria do gráfico.

```
poluicao %>%
  ggplot(aes(temp2)) +
    geom_histogram(
      bins = 10,
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.55: Distribuição da temperatura",
      x = "Temperatura",
      y = "Frequência"
    ) +
    tema
```

Figura 3.55: Distribuição da temperatura



Exercício 3.8

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

Solução. Carregando os dados:

```
salarios <- readxl::read_xls(paste0(data_dir, "salarios.xls"), skip = 4)
```

O conjunto de dados contém valores de salários para quatro profissões (Profissional de Segurança, Mecânico, Administrador e Engenheiro Eletricista) em trinta cidades ao redor do mundo.

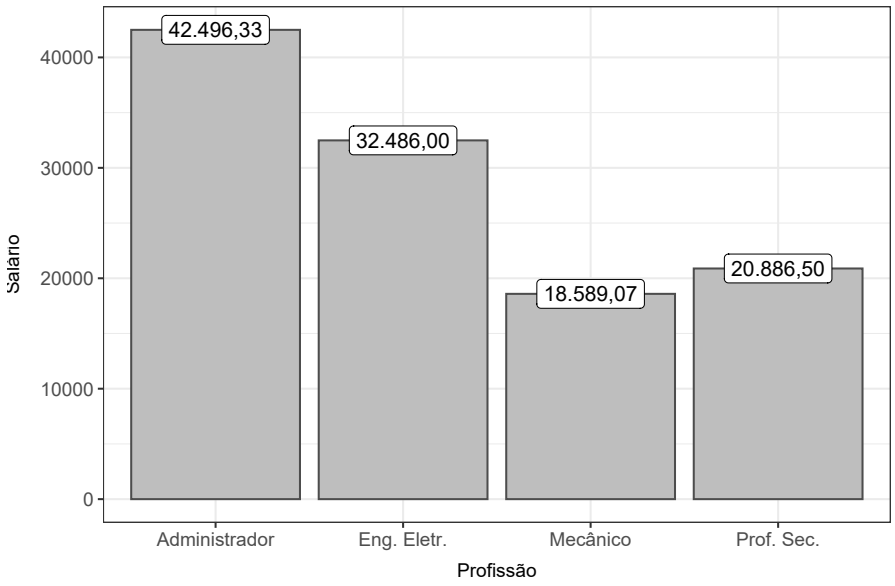
Vamos fazer uma transformação nos dados para facilitar a visualização:

```
salario2 <- salarios %>%
              gather(key = "Profissao", value = "Salario", -"Cidade")
```

Agora vamos visualizar os salários por profissão:

```
salario2 %>%
  group_by(Profissao) %>%
  summarise(
    Media = mean(Salario)
  ) %>%
  ggplot(aes(x = Profissao, y = Media)) +
    geom_col(
      fill = "grey",
      color = "grey30"
    ) +
    geom_label(aes(x = Profissao, y = Media, label = format(round(Media, 2), decimal.mark = ",", big.mark = ".")
    labs(
      title = "Figura 3.56: Salário médio por profissão",
      x = "Profissão",
      y = "Salário"
    ) +
    tema
```

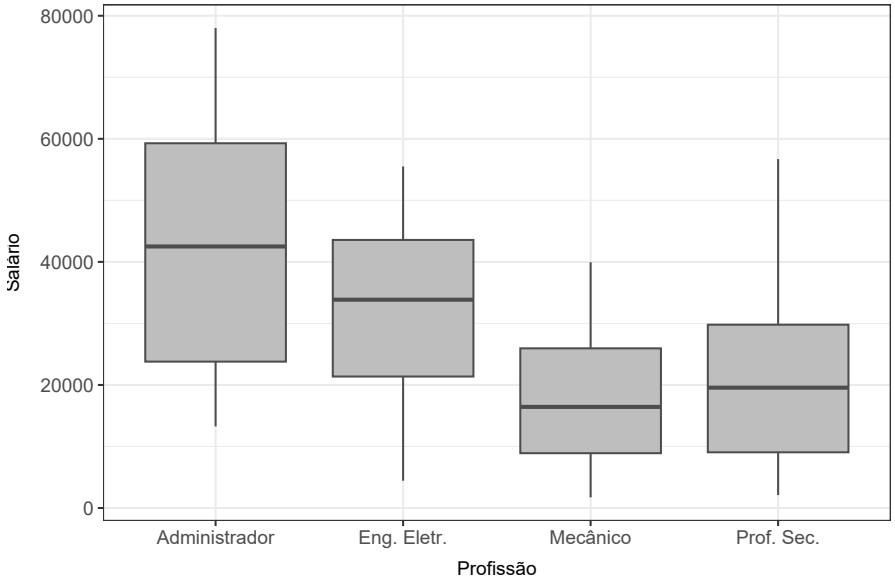
Figura 3.56: Salário médio por profissão




```
salario2 %>%
  ggplot(aes(Profissao, Salario, group = Profissao)) +
    geom_boxplot(
      fill = "grey",
```

```
    color = "grey30",
  ) +
  labs(
    title = "Figura 3.57: Distribuição dos salários por profissão",
    x = "Profissão",
    y = "Salário"
  ) +
  tema
```

Figura 3.57: Distribuição dos salários por profissão



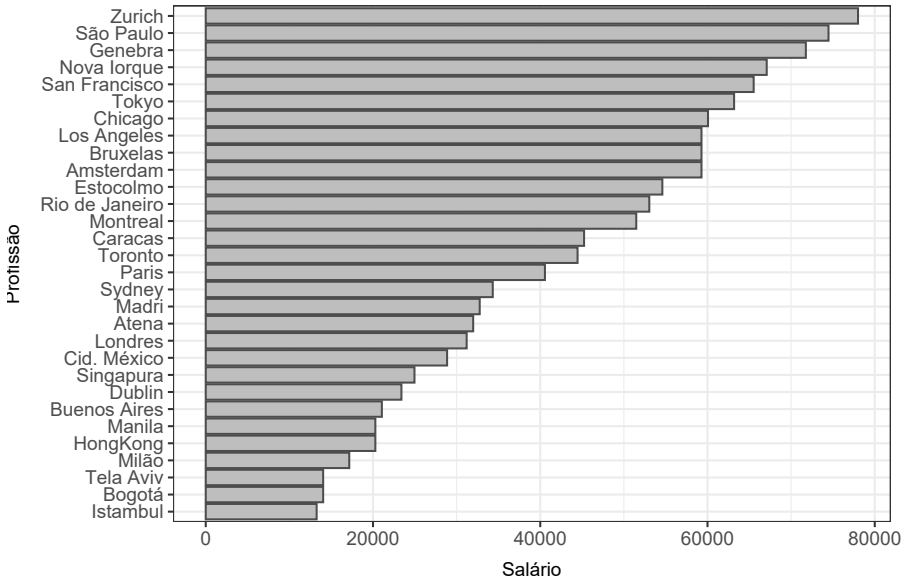
A média geral dos salários é 28.614,47 e o desvio padrão é 17.974,29.

Agora vamos focar apenas nos administradores. A distribuição dos salários por cidade é mostrada no gráfico a seguir:

```
order <- arrange(salarios, Administrador)$Cidade

salarios %>%
  ggplot(aes(Cidade, Administrador)) +
    geom_col(
      fill = "grey",
      color = "grey30",
    ) +
    coord_flip() +
```

```
  scale_x_discrete(limits = order) +
  labs(
    title = "Figura 3.58: Distribuição dos salários por profissão",
    x = "Profissão",
    y = "Salário"
  ) +
  tema
```

Figura 3.58: Distribuição dos salários por profissão



A média salarial dos administradores é 42.496,33, o desvio padrão é 20.366,78 e a mediana é 42.510.

Exercício 3.9

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

Solução. Carregamos os dados com a função `readxls` do pacote **readxl**:

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

```
## New names:
## * `Ano` -> `Ano...1`
```

```
## * `Chuva` -> `Chuva...2`
## * `Ano` -> `Ano...3`
## * `Chuva` -> `Chuva...4`
## * `Ano` -> `Ano...5`
## * `Chuva` -> `Chuva...6`
## * `Ano` -> `Ano...7`
## * `Chuva` -> `Chuva...8`
```

Como os dados estão mal organizados na planilha, precisamos fazer um ajuste.

```
a <- precipitacao %>%
        select(Ano...1, Chuva...2)

b <- precipitacao %>%
        select(Ano...3, Chuva...4)

c <- precipitacao %>%
        select(Ano...5, Chuva...6)

d <- precipitacao %>%
        select(Ano...7, Chuva...8)

precipitacao <- bind_rows(a, b, c, d) %>% drop_na()
```

```
## New names:
## New names:
## New names:
## New names:
## * `Ano...1` -> `Ano`
## * `Chuva...2` -> `Chuva`
```

Utilizamos a função `stem()` para construir o diagrama de ramos e folhas:

```
stem(precipitacao$Chuva)
```

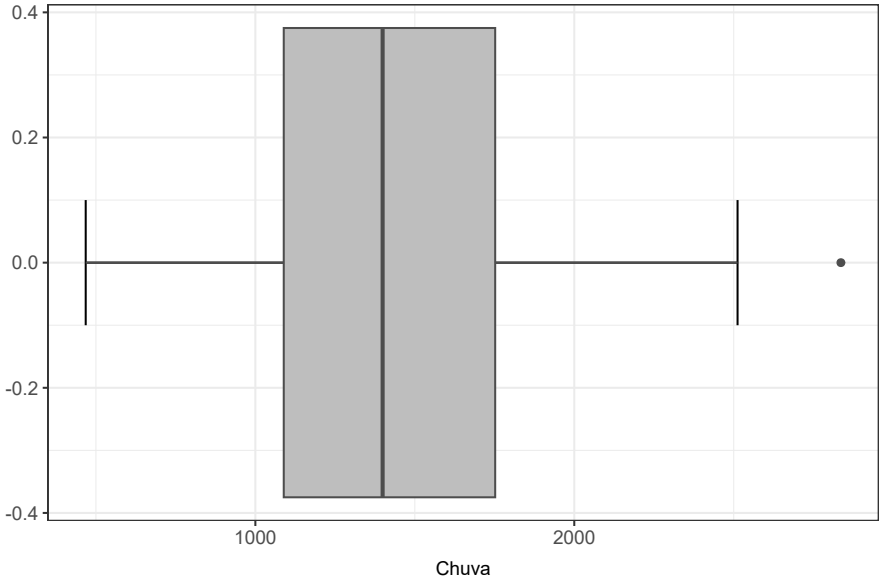
```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##    4 | 7
##    5 | 003
##    6 | 06
```

```
##    7 | 045889
##    8 | 12335688
##    9 | 24468
##   10 | 00123445578999
##   11 | 0113444599
##   12 | 0123334566799
##   13 | 011223667889
##   14 | 02233555667799
##   15 | 111345577899
##   16 | 0336
##   17 | 1223345578
##   18 | 1144556689
##   19 | 114
##   20 | 035689
##   21 | 044
##   22 | 6
##   23 | 38
##   24 | 135689
##   25 | 011
##   26 |
##   27 |
##   28 | 4
```

Utilizamos **ggplot2** para construir o boxplot.

```
precipitacao %>%
  ggplot(aes(Chuva)) +
    stat_boxplot(geom = "errorbar", width = 0.2) +
    geom_boxplot(
      fill = "grey",
      color = "grey30",
    ) +
    labs(
      title = "Figura 3.59: Precipitação atmosférica em Fortaleza-CE",
      x = "Chuva"
    ) +
    tema
```

Figura 3.59: Precipitação atmosférica em Fortaleza-CE



Salvamos o conjunto de dados para futuras análises:

```
write_csv(precipitacao, paste0(data_dir, "precipitacao.csv"))
```

Exercício 3.10

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

Solução. Carregamos os dados com **readxl**:

```
manchas <- readxl::read_xls(paste0(data_dir, "manchas.xls"), skip = 4)
```

```
## New names:
## * `Ano` -> `Ano...1`
## * `Número` -> `Número...2`
## * `Ano` -> `Ano...3`
## * `Número` -> `Número...4`
## * `Ano` -> `Ano...5`
## * `Número` -> `Número...6`
## * `Ano` -> `Ano...7`
## * `Número` -> `Número...8`
```

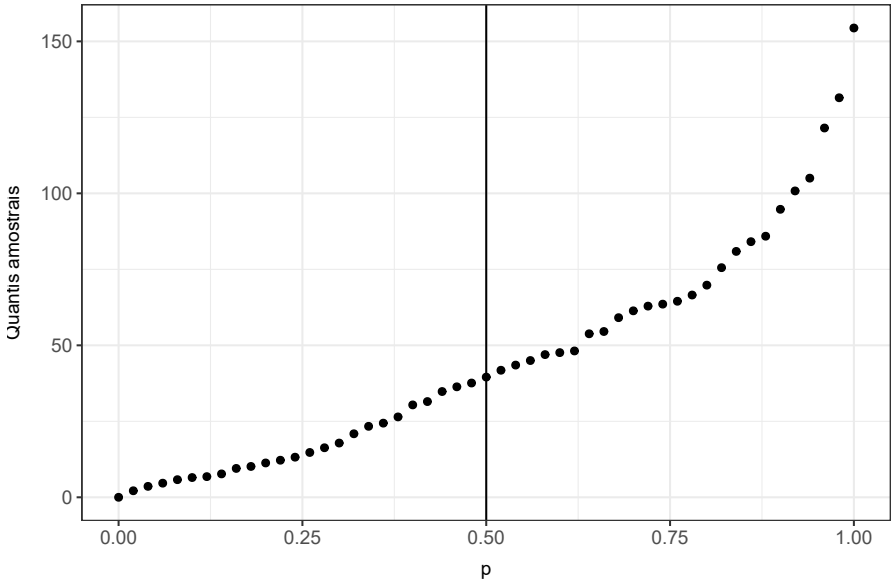
```
a <- manchas %>% select(Ano...1, Número...2)
b <- manchas %>% select(Ano...3, Número...4)
c <- manchas %>% select(Ano...5, Número...6)
d <- manchas %>% select(Ano...7, Número...8)

manchas <- bind_rows(a, b, c, d) %>% drop_na()
```

```
## New names:
## New names:
## New names:
## New names:
## * `Ano...1` -> `Ano`
## * `Número...2` -> `Número`
```

Agora vamos construir o gráfico que quantis:

```
p <- seq(0, 1, 0.02)
q <- quantile(manchas$Número, p)

data.frame(p, q) %>%
  ggplot(aes(p, q)) +
    geom_point() +
    geom_vline(aes(xintercept = 0.5)) +
    labs(
      title = "Figura 3.60: Quantis para o número de manchas solares",
      x = "p",
      y = "Quantis amostrais"
    ) +
    tema
```

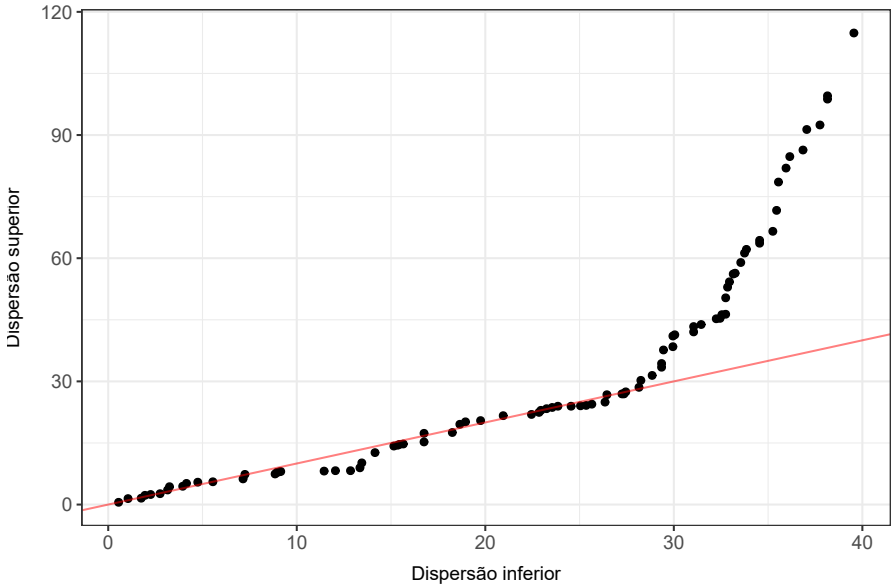
Figura 3.60: Quantis para o número de manchas solares



Para o gráfico de simetria (dispersão), temos o seguinte:

```
manchas %>%
  ggplot() +
    geom_symmetry(manchas$Número) +
    geom_abline(aes(intercept = 0, slope = 1), color = "red", alpha = .5) +
    labs(
      title = "Figura 3.61: Gráfico de simetria para o número de manchas solares",
      x = "Dispersão inferior",
      y = "Dispersão superior"
    ) +
    tema
```

Figura 3.61: Gráfico de simetria para o número de manchas solares



Os gráficos acima nos dão uma indicação de vele assimetria à direita. O que pode ser verificado com o Coef. de Assimetria de Fisher-Pearson Ajustado 0.8596009.

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 iual a zero no caso de distribuições simétricas. Calcule-a para os dados do Exercício 3.4.

Solução. Primeiro vamos definir uma função para calcular a medida.

```
assimetria <- function(x) {
  q <- quantile(x, c(0.25, 0.5, 0.75))

  ((q[[3]] - q[[2]]) - (q[[2]] - q[[1]]))/(q[[3]] - q[[1]])
}
```

Usando essa função, temos que o coeficiente de assimetria para a variável `vt` do conjunto `vento.xls` é de 0.0788382.

Exercício 3.12

Os dados disponíveis no arquivo `endometrione.xls` [Abrão et al., 1997] são provenientes de um estudo em que o objetivo é verificar se existe diferença entre os grupos de doentes e controles quanto a algumas caracteristicas observadas.

a) 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 capitulo? Faça-a e tire suas conclusões.
b) Compare as distribuições das variáveis idade e concentração de PCR durante a menstruação (`PCRa`) para pacientes dos grupos controle e doente utilizando medidas resumo (minimo, máximo, quartis, mediana, média, desvio padrão, etc.), *boxplots*, histogramas, gráficos de médias e gráficos QQ. Como você considerou os valores $< 0,5$ da variável `PCRa` nesses cálculos? Você sugeriria uma outra maneira para considerar tais valores?
c) 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. Como sempre, vamos iniciar com o carregamento e ajuste nos dados:

```
endometriose <- readxl::read_xls(paste0(data_dir, "endometriose.xls"), sheet = "dados")

head(endometriose, 10)
```

```
## # A tibble: 10 x 13
##    Grupo   Paciente Idade Gestação Partos Abortos Dismenorréia Dispareunia  AFSr
##    <chr>      <dbl> <dbl>    <dbl>  <dbl>   <dbl> <chr>        <chr>       <dbl>
##  1 Contro~        1    26        3      3       0 L            N               0
##  2 Contro~        2    37        4      3       1 N            P               0
##  3 Contro~        3    37        4      4       0 N            N               0
##  4 Contro~        4    35        3      3       0 L            N               0
##  5 Contro~        5    34        4      3       1 N            N               0
##  6 Contro~        6    38        5      5       0 L            N               0
##  7 Contro~        7    30        5      4       1 N            N               0
##  8 Contro~        8    38       11      7       4 N            N               0
##  9 Contro~        9    36        7      6       1 N            N               0
## 10 Contro~       10    41        4      3       1 N            N               0
## # i 4 more variables: `CA125/A` <chr>, `CA125/B` <chr>, PCRa <chr>, PCRb <chr>
```

Vamos transformar em fator as variáveis `Grupo`, `Dismenorréia` e `Dispareunia`. Vamos também ubstituir o valor `<0,5` por zero nas variáveis `CA125/A`, `CA125/B`, `PCRa` e `PCRb`.

Por fim, transformaremos em inteiras as variáveis `Idade`, `Gestação`, `Partos`, `Abortos` e `AFSr`.

```
dismenorreia_levels <- c("Não tem", "Leve", "Moderada", "Intensa", "Sem informação")
dismenorreia <- function(x) {
  x %>%
    str_replace("S/", "Sem informação") %>%
    str_replace("N", "Não tem") %>%
    str_replace("L", "Leve") %>%
    str_replace("M", "Moderada") %>%
    str_replace("I", "Intensa")
}

dispareunia <- function(x) {
  x %>%
    str_replace("\\.", "Sem informação") %>%
    str_replace("N", "Não tem") %>%
    str_replace("P\\b", "Penetração") %>%
    str_replace("PRO", "Profunda") %>%
    str_replace("2", "Penetração e Profunda")
}

zerar <- function(x) {
  x %>%
    str_replace("<0,5", "0")
}

endometriose <- endometriose %>%
                  mutate(
                    Grupo = parse_factor(Grupo),
                    Dismenorréia = parse_factor(dismenorreia(Dismenorréia), levels = dismenorreia_levels, ordere
                    Dispareunia = parse_factor(dispareunia(Dispareunia)),
                    `CA125/A` = parse_double(zerar(`CA125/A`)),
                    `CA125/B` = parse_double(zerar(`CA125/B`)),
                    PCRa = parse_double(zerar(PCRa)),
                    PCRb = parse_double(zerar(PCRb)),
                    Idade = as.integer(Idade),
                    Gestação = as.integer(Gestação),
                    Partos = as.integer(Partos),
                    Abortos = as.integer(Abortos),
                    AFSr = as.integer(AFSr)
                  )

statds_write(endometriose, "endometriose.csv")
```

Pacientes doentes apresentam mais dor na menstruação do que as pacientes não doentes?

Em primeiro lugar, vamos construir uma tabela de contigência com os graus de dismenorréia entre os grupos.

```
endometriose %>%
  group_by(Dismenorréia, Grupo) %>%
  count() %>%
  spread(Grupo, n) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.7:** Distribuição da ocorrência de dismenorréia entre casos e controles",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ","),
  )
```

Tabela 3.7: Tabela 3.7: Distribuição da ocorrência de dismenorréia entre casos e controles

Dismenorréia	Controle	Doente
Não tem	10	-
Leve	4	3
Moderada	1	14
Intensa	-	17
Sem informação	-	1

Através da Tabela 3.7, é possível perceber que 88,57% das pacientes com endometriose apresentam dor moderada ou intensa, contra apenas 6,67% das pacientes que não apresentam esta condição. Ou seja, as pacientes com a doença apresentam mais dor na mestruação do que as demais.

Distribuição das variáveis `Idade` e `PCRa` entre os grupos

Primeiro vamos calcular as medidas globais de resumo para cada uma das variáveis.

```
endometriose %>%
  select(Idade, PCRa) %>%
```

```
  map(summ) %>%
  as.data.frame() %>%
  t() %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.8:** Medidas de resumo para as variáveis `Idade` e `PCRa`",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.8: Tabela 3.8: Medidas de resumo para as variáveis `Idade` e `PCRa`

	n	Mínimo	Q1	Mediana	Q3	Máximo	Média	Desvio Padrão	Distância Interquartil
Idade	50	20	27,00	32,00	37,0	43	32,04	5,76	10,00
PCRa	50	0	1,27	3,55	6,9	32	6,33	8,06	5,62

Para o grupo de controle, temos o seguinte:

```
endometriose %>%
  filter(Grupo == "Controle") %>%
  select(Idade, PCRa) %>%
  map(summ) %>%
  as.data.frame() %>%
  t() %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.9:** Medidas de resumo para o grupo de controle",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.9: Tabela 3.9: Medidas de resumo para o grupo de controle

	n	Mínimo	Q1	Mediana	Q3	Máximo	Média	Desvio Padrão	Distância Interquartil
Idade	15	26	33	36,0	37,5	41,0	35,13	3,80	4,5
PCRa	15	0	0	1,2	2,3	3,8	1,35	1,44	2,3

Para o grupo de casos, obtemos as seguintes medidas:

```
endometriose %>%
  filter(Grupo == "Doente") %>%
  select(Idade, PCRa) %>%
  map(summ) %>%
  as.data.frame() %>%
  t() %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.10:** Medidas de resumo para o grupo de casos",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.10: Tabela 3.10: Medidas de resumo para o grupo de casos

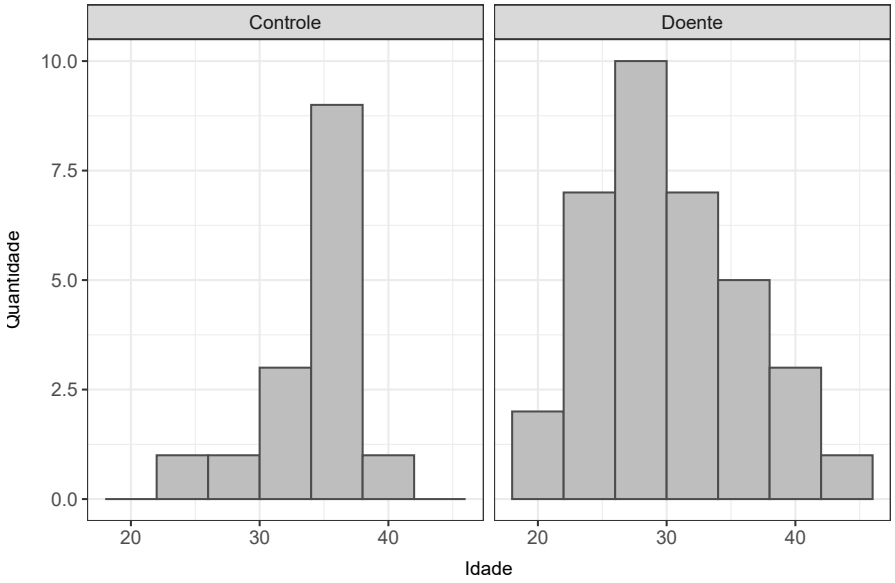
	n	Mínimo	Q1	Mediana	Q3	Máximo	Média	Desvio Padrão	Distância Interquartil
Idade	35	20	26,5	29,0	35,5	43	30,71	5,98	9,0
PCRa	35	0	3,2	4,6	12,5	32	8,46	8,79	9,3

Vamos também construir alguns gráficos para visualizar melhor os dados:

```
endometriose %>%
  ggplot(aes(Idade)) +
    geom_histogram(
      binwidth = 4,
      fill = "grey",
      color = "grey30"
```

```
  ) +
  facet_wrap(~Grupo) +
  labs(
    title = "Figura 3.62: Distribuição da idade entre pacientes",
    x = "Idade",
    y = "Quantidade"
  ) +
  tema
```

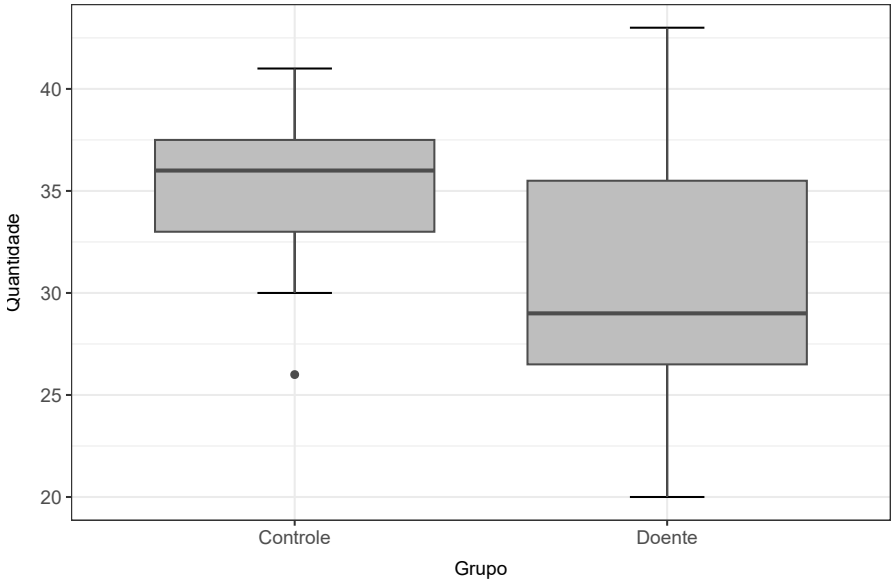
Figura 3.62: Distribuição da idade entre pacientes




```
endometriose %>%
  ggplot(aes(Grupo, Idade)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = "Figura 3.63: Distribuição da idade entre pacientes",
      y = "Quantidade"
```

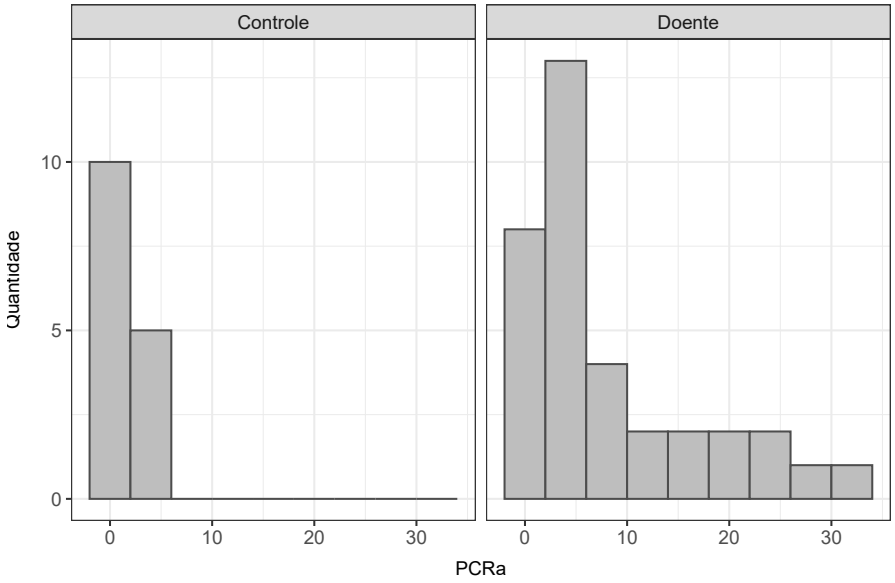
```
  ) +
  tema
```

Figura 3.63: Distribuição da idade entre pacientes



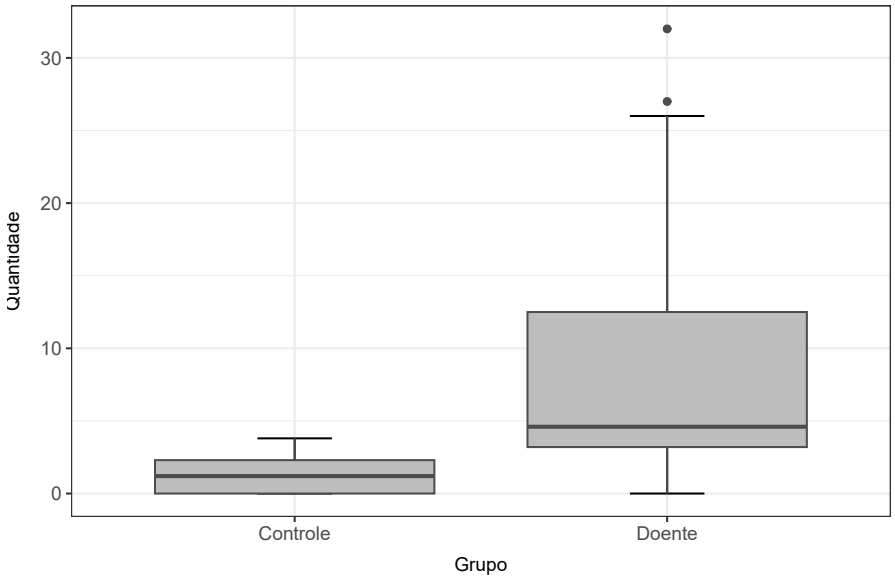

```
endometriose %>%
  ggplot(aes(PCRa)) +
    geom_histogram(
      binwidth = 4,
      fill = "grey",
      color = "grey30"
    ) +
    facet_wrap(~Grupo) +
    labs(
      title = "Figura 3.64: Distribuição da variável `PCRa` entre pacientes",
      x = "PCRa",
      y = "Quantidade"
    ) +
    tema
```

Figura 3.64: Distribuição da variável `PCRa` entre pacientes




```
endometriose %>%
  ggplot(aes(Grupo, PCRa)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = bquote(bold("Figura 3.65:")~"Distribuição da variável `PCRa` entre pacientes"),
      y = "Quantidade"
    ) +
    tema
```

Figura 3.65: Distribuição da variável `PCRa` entre pacientes



Notamos com esses dados que a idade das pacientes com endrometriose varia de 20 a 43 anos, com média de 30,71 anos e desvio padrão de 5,98 anos. Já entre o grupo de controle, a variação de idade é menos (entre 26 e 41 anos), com média de 35,13 anos e desvio padrão de 3,80 anos.

Para a variável PCRa, temos uma discrepância considerável. Mulheres com endometriose tem uma concentração de mais de 6 vezes maior de PCR durante a menstruação do que as pacientes do grupo de controle.

Optamos por zerar o valor das observações cuja medida foi $ < 0,5$. Uma vez que 0,5 é o menor valor que o instrumento de aferição consegue detectar, entendemos ser justo considerar essas medidas como nula. Seria uma opção desconsiderar essas medidas? Cabe avaliar.

Comparando o número de gestações entre os grupos

```
endometriose %>%
  group_by(Grupo, Gestação) %>%
  count() %>%
  pivot_wider(names_from = Grupo, values_from = n, values_fill = 0) %>%
  arrange(Gestação)
```

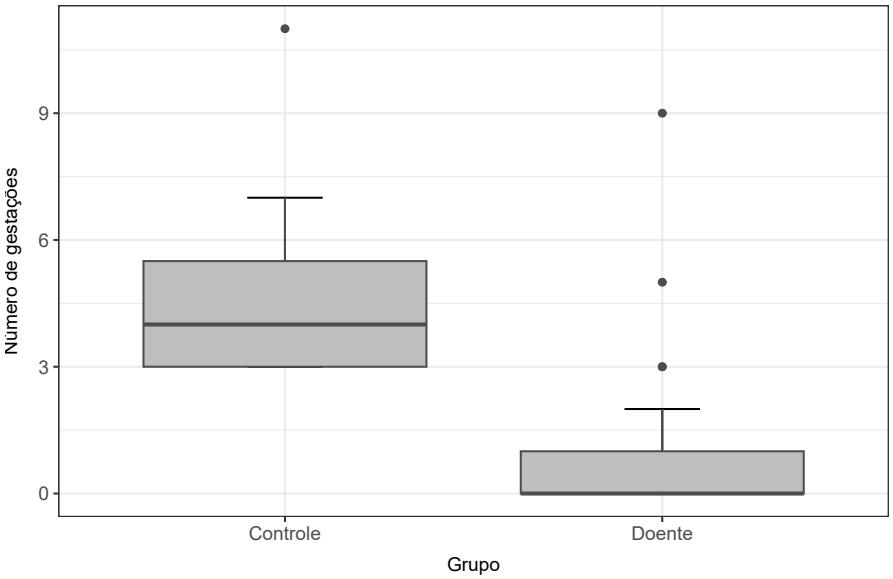
```
## # A tibble: 10 x 3
```

```
## # Groups:   Gestação [10]
##    Gestação Controle Doente
##       <int>    <int>  <int>
##  1        0        0     18
##  2        1        0      9
##  3        2        0      4
##  4        3        5      2
##  5        4        4      0
##  6        5        2      1
##  7        6        1      0
##  8        7        2      0
##  9        9        0      1
## 10       11        1      0
```

Vamos também representar graficamente a relação entre os grupos usando um *box-plot*.

```
endometriose %>%
  ggplot(aes(x = Grupo, y = Gestação)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = bquote(bold("Figura 3.66:")~"Número de gestações entre mulheres com e sem endometriose"),
      x = "Grupo",
      y = "Número de gestações"
    ) +
    tema
```

Figura 3.66: Número de gestações entre mulheres com e sem endometriose



Notamos que o número de gestações entre mulheres com endometriose é muito menor do que entre as mulheres do grupo controle. Uma parte desse efeito poderia ser explicado pela idade das pacientes (a média de idade das pacientes com endometriose é menor do que as do grupo de controle), mas certamente a doença em si é o maior fator pra o menor número de gestações.

Exercício 3.13

Os dados encontrados no arquivo `enforco.xls` [Braga, 1998] 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 clinicas dos pacientes e foram registras uma única vez.

a) Descreva a distribuição da variável consumo de oxigênio (`VO2`) 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.
b) Descreva a distribuição da classe funcional NYHA por meio de uma tabela de frequências. Utilize um método gráfico para representar essa tabela.

Solução.

Preparação dos dados

Iniciaremos carregando os dados e ajustando os nomes das colunas.

```
esforco <- readxl::read_xls(paste0(data_dir, "esforco.xls"), sheet = "dados", skip = 3, col_names = FALSE)
colnames(esforco) <- c(
                       "id_grupo", "iniciais", "etiologia", "sexo", "data_espirometrico",
                       "idade", "altura", "peso", "superficie_corporal", "imc",
                       "nyha", "weber", "fc_rep", "vo2_rep", "rer_rep", "vo2_fc_rep",
                       "ve_vo2_rep", "ve_vco2_rep", "aumento_carga_rep", "teste_maximo_rep",
                       "carga_lan", "perc_max_carga_lan", "fc_lan", "vo2_lan",
                       "perc_max_vo2_lan", "rer_lan", "perc_max_rer_lan", "vo2_fc_lan",
                       "perc_max_vo2_fc_lan", "ve_vo2_lan", "perc_max_ve_vo2_lan",
                       "ve_vco2_lan", "perc_max_ve_vco2_lan", "carga_pcr", "perc_max_carga_pcr",
                       "fc_pcr", "vo2_pcr", "perc_max_vo2_pcr", "rer_pcr", "perc_max_rer_pcr",
                       "vo2_fc_pcr", "perc_max_vo2_fc_pcr", "ve_vo2_pcr", "perc_max_ve_vo2_pcr",
                       "ve_vco2_pcr", "perc_max_ve_vco2_pcr", "carga_w", "fc_w",
                       "vo2_w", "rer_w", "vo2_fc_w", "ve_vo2_w", "ve_vco2_w",
                       "tempo_rampa", "vo2_wr_slope", "ve_vco2_slope", "data_obito",
                       "data_ultima_obs", "cirurgia"
                     )
```

Vamos também criar algumas funções que nos ajudarão a ajustar os dados.

```
map_etiologia <- function(x) {
  x %>%
    str_replace("CH\\b", "Chagasíaco") %>%
    str_replace("ID\\b", "Idiopático") %>%
    str_replace("IS\\b", "Isquêmico") %>%
    str_replace("C\\b", "Controle")
}

map_sexo <- function(x) {
  x %>%
    str_replace("M", "Masculino") %>%
    str_replace("F", "Feminino")
}

map_sim_nao_na <- function(x) {
  x %>%
    str_replace("S", "Sim") %>%
    str_replace("N", "Não") %>%
```

```
    str_replace("\\.", NA_character_)
}
```

Agora vamos aos ajustes:

```
esforco <- esforco %>%
              mutate(
                id = ifelse(etiologia == 'C', id_grupo + 87, id_grupo),
                grupo = parse_factor(ifelse(etiologia == 'C', "Controle", "Caso")),
                id_grupo = as.integer(id_grupo),
                etiologia = parse_factor(map_etiologia(etiologia)),
                sexo = parse_factor(map_sexo(sexo)),
                data_espirometrico = dmy(data_espirometrico),
                idade = as.integer(idade),
                altura = as.integer(altura),
                nyha = parse_factor(as.character(nyha), levels = c("1", "2", "3", "4"), ordered = TRUE, na = "0"
                weber = parse_factor(weber, levels = c("A", "B", "C", "D", "E", NA), ordered = TRUE),
                teste_maximo_rep = parse_factor(map_sim_nao_na(teste_maximo_rep)),
                carga_lan = parse_integer(carga_lan, na = "."),
                perc_max_carga_lan = parse_double(perc_max_carga_lan, na = "."),
                fc_lan = parse_double(fc_lan, na = "."),
                vo2_lan = parse_double(vo2_lan, na = "."),
                perc_max_vo2_lan = parse_double(perc_max_vo2_lan, na = "."),
                rer_lan = parse_double(rer_lan, na = "."),
                perc_max_rer_lan = parse_double(perc_max_rer_lan, na = "."),
                vo2_fc_lan = parse_double(vo2_fc_lan, na = "."),
                perc_max_vo2_fc_lan = parse_double(perc_max_vo2_fc_lan, na = "."),
                ve_vo2_lan = parse_double(ve_vo2_lan, na = "."),
                perc_max_ve_vo2_lan = parse_double(perc_max_ve_vo2_lan, na = "."),
                ve_vco2_lan = parse_double(ve_vco2_lan, na = "."),
                perc_max_ve_vco2_lan = parse_double(perc_max_ve_vco2_lan, na = "."),
                carga_pcr = parse_double(carga_pcr, na = "."),
                perc_max_carga_pcr = parse_double(perc_max_carga_pcr, na = "."),
                fc_pcr = parse_double(fc_pcr, na = "."),
                vo2_pcr = parse_double(vo2_pcr, na = "."),
                perc_max_vo2_pcr = parse_double(perc_max_vo2_pcr, na = "."),
                rer_pcr = parse_double(rer_pcr, na = "."),
                perc_max_rer_pcr = parse_double(perc_max_rer_pcr, na = "."),
                vo2_fc_pcr = parse_double(vo2_fc_pcr, na = "."),
                perc_max_vo2_fc_pcr = parse_double(perc_max_vo2_fc_pcr, na = "."),
                ve_vo2_pcr = parse_double(ve_vo2_pcr, na = "."),
                perc_max_ve_vo2_pcr = parse_double(perc_max_ve_vo2_pcr, na = "."),
                ve_vco2_pcr = parse_double(ve_vco2_pcr, na = "."),
```

```
                perc_max_ve_vco2_pcr = parse_double(perc_max_ve_vco2_pcr, na = "."),
                carga_w = parse_double(carga_w, na = "."),
                tempo_rampa  = parse_double(tempo_rampa, na = "."),
                vo2_wr_slope  = parse_double(vo2_wr_slope, na = "."),
                data_obito = dmy(data_obito),
                data_ultima_obs = dmy(data_ultima_obs)
              )
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `data_obito = dmy(data_obito)`.
## Caused by warning:
## !  47 failed to parse.
```

Com o arquivo organizado, vamos salvá-lo para uso futuro como `esforco_completo.csv`.

```
write_csv(esforco, paste0(data_dir, "esforco_completo.csv"))
```

Consumo de oxigênio nos quatro momentos

Inicialmente precisamos destacar que as variáveis de interesse são `vo2_rep`, `vo2_lan`, `vo2_pcr` e `vo2_w`. Vamos isolar essas variáveis e utilizar a função `gather()` para reunir os dados dos momentos em um formato mais amigável.

```
vo2 <- esforco %>%
          select(id, grupo, vo2_rep, vo2_lan, vo2_pcr, vo2_w) %>%
          gather(vo2_rep, vo2_lan, vo2_pcr, vo2_w, key = "Momento", value = "vo2") %>%
          mutate (
            Momento = Momento %>%
                        str_replace("vo2_rep", "Repouso") %>%
                        str_replace("vo2_lan", "Limiar anaeróbico") %>%
                        str_replace("vo2_pcr", "Ponto de Compensação") %>%
                        str_replace("vo2_w", "Pico")
          )

head(vo2)
```

```
## # A tibble: 6 x 4
##      id grupo Momento   vo2
##   <dbl> <fct> <chr>   <dbl>
```

```
## 1     1 Caso  Repouso   5.9
## 2     2 Caso  Repouso   3.4
## 3     3 Caso  Repouso   3
## 4     4 Caso  Repouso   3.8
## 5     5 Caso  Repouso   3.2
## 6     6 Caso  Repouso   3.8
```

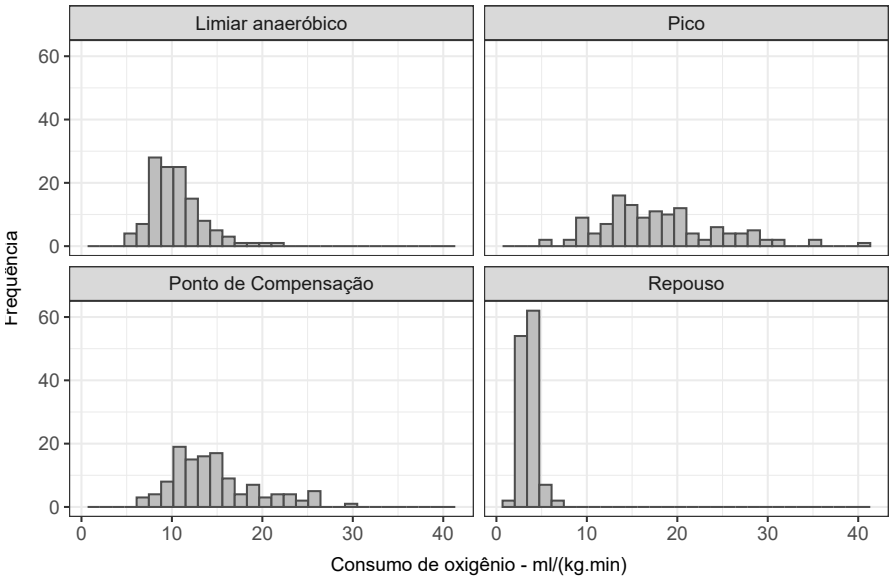
Agora vamos avaliar a disbtibuição da variável `vo2` em cada um dos momentos:

```
vo2 %>%
  ggplot(aes(vo2)) +
    geom_histogram(
      fill = "grey",
      color = "grey30"
    ) +
    facet_wrap(~Momento) +
    labs(
      title = bquote(bold("Figura 3.67:")~"Consumo de oxigênio durante teste de esforço"),
      x = "Consumo de oxigênio - ml/(kg.min)",
      y = "Frequência"
    ) +
    tema
```

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

```
## Warning: Removed 9 rows containing non-finite values (`stat_bin()`).
```

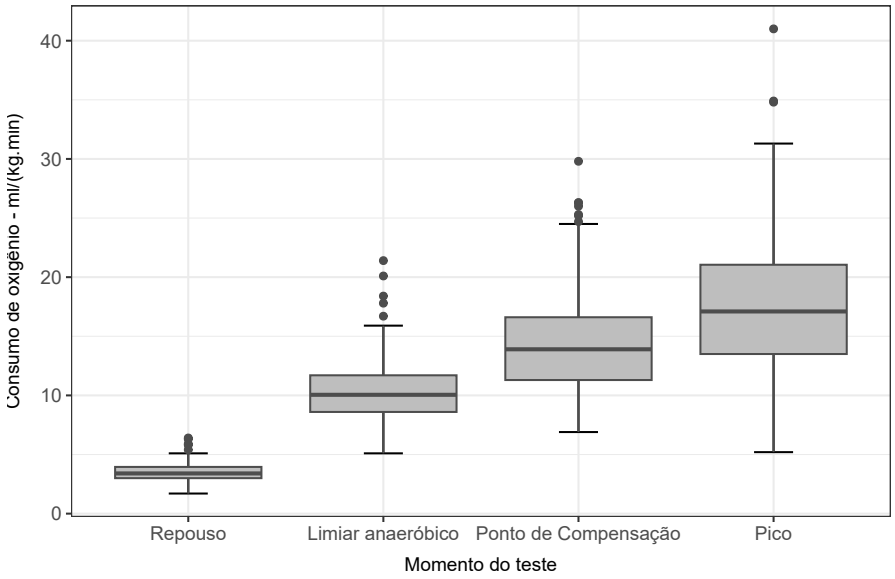
Figura 3.67: Consumo de oxigênio durante teste de esforço




```
vo2 %>%
  ggplot(aes(
      x = factor(Momento, level = c("Repouso", "Limiar anaeróbico", "Ponto de Compensação", "Pico")),
      y = vo2
  )) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = bquote(bold("Figura 3.68:")~"Consumo de oxigênio durante teste de esforço"),
      x = "Momento do teste",
      y = "Consumo de oxigênio - ml/(kg.min)"
    ) +
    tema
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_boxplot()`).
## Removed 9 rows containing non-finite values (`stat_boxplot()`).
```

Figura 3.68: Consumo de oxigênio durante teste de esforço



Vamos calcular também a média e desvio padrão para o consumo de oxigênio em cada momento:

```
resumo <- vo2 %>%
            group_by(Momento, grupo) %>%
            summarise(
              `Média` = mean(vo2, na.rm = TRUE),
              `Desvio Padrão` = sd(vo2, na.rm = TRUE),
              Q1 = quantile(vo2, 0.25, na.rm = TRUE)[[1]],
              `Mediana` = median(vo2, na.rm = TRUE),
              Q3 = quantile(vo2, 0.75, na.rm = TRUE)[[1]],
              `Distância Interquartil` = IQR(vo2, na.rm = TRUE)
            ) %>%
            arrange(factor(Momento, level = c("Repouso", "Limiar anaeróbico", "Ponto de Compensação", "Pico")))
```

```
## `summarise()` has grouped output by 'Momento'. You can override using the
## `.groups` argument.
```

```
resumo %>%
  kable(
    format = "pipe",
```

```
    caption = "**Tabela 3.11:** Consumo de oxigênio por momento de avaliação",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.11: Tabela 3.11: Consumo de oxigênio por momento de avaliação

Momento	grupo	Média	Desvio Padrão	Q1	Mediana	Q3	Distância Interquartil
Repouso	Caso	3,58	0,91	3,00	3,40	3,95	0,95
Repouso	Controle	3,50	0,60	3,00	3,45	3,85	0,85
Limiar anaeróbico	Caso	9,98	2,73	8,47	9,70	11,03	2,55
Limiar anaeróbico	Controle	11,53	2,84	9,60	10,85	13,77	4,17
Ponto de Compensação	Caso	13,32	4,05	10,60	12,80	15,10	4,50
Ponto de Compensação	Controle	17,68	4,67	13,83	16,15	21,50	7,67
Pico	Caso	15,78	5,43	12,15	14,50	18,45	6,30
Pico	Controle	23,00	5,82	18,28	22,75	27,72	9,45

Os gráficos nos mostram que alguns pacientes possuem um consumo excessivamente alto de oxigênio. Apesar desses valores, os dados parecem apresentar tendência normal. Notamos ainda uma menor consumo de oxigênio para o grupo de pacientes com alguma doença pré-existente.

Abaixo, listamos os identificadores dos valores discrepantes.

```
resumo2 <- resumo %>%
            transmute(
              m = Momento,
              l = Q3 + 1.5 * `Distância Interquartil`
            )

vo2 %>%
  left_join(resumo2, by = join_by(Momento == m)) %>%
  filter(vo2 > l) %>%
  select(id) %>%
```

```
  distinct() %>%
  array()
```

```
## Warning in left_join(., resumo2, by = join_by(Momento == m)): Detected an unexpected many-
to-many relationship between `x` and `y`.
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 1 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
## [[1]]
##  [1]   1  29  32  52  85  34  38  42  45  95  96 103 106 109 122 123  18  91 104
## [20] 111 115 126 101 110 127
```

Observação

Para melhorar a análise, poderíamos montar gráficos separados para casos e controles. Isso facilitaria a comparação entre os dois grupos em cada etapa.

Exercício 3.14

Na tabela abaixo estão indicadas as durações de 335 lâmpadas.

Tabela 3.12: Tabela 3.12: Duração de 335 lâmpadas

Duração (horas)	Número de lâmpadas
$0 \vdash 100$	82
$100 \vdash 200$	71
$200 \vdash 300$	68
$300 \vdash 400$	56
$400 \vdash 500$	43
$500 \vdash 800$	15

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

Solução. Vamos começar colocando os dados da tabela em um data frame.

```
lampadas <- tribble(
  ~min, ~max, ~freq,
  0, 99, 82,
  100, 199, 71,
  200, 299, 68,
  300, 399, 56,
  400, 499, 43,
  500, 599, 15
)
```

Vamos agora adicionar a frequência relativa e a frequência acumulada ao nosso data frame:

```
lampadas <- lampadas %>%
                mutate(
                  freq_acu = cumsum(freq),
                  freq_rel = freq/sum(freq),
                  freq_rel_acu = cumsum(freq_rel)
                )

lampadas
```

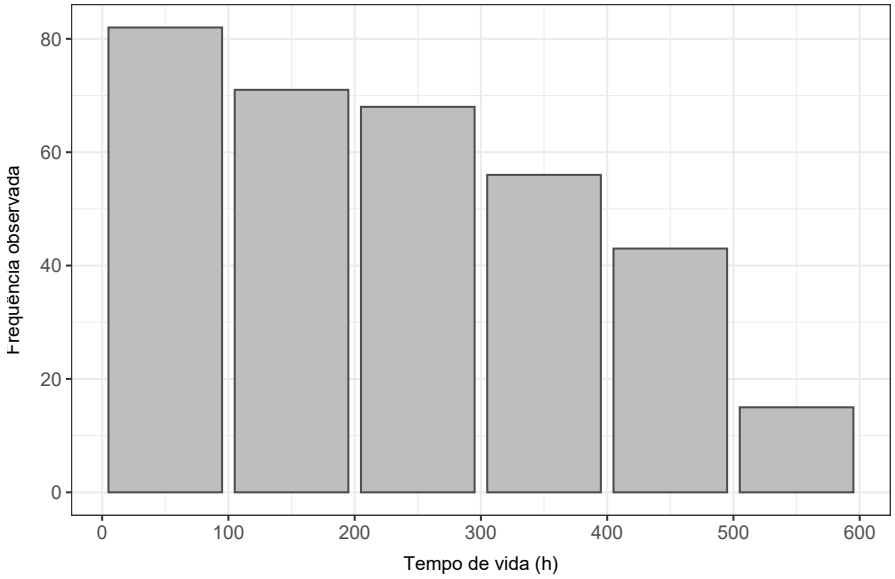
```
## # A tibble: 6 x 6
##     min   max  freq freq_acu freq_rel freq_rel_acu
##   <dbl> <dbl> <dbl>    <dbl>    <dbl>        <dbl>
## 1     0    99    82       82   0.245         0.245
## 2   100   199    71      153   0.212         0.457
## 3   200   299    68      221   0.203         0.660
## 4   300   399    56      277   0.167         0.827
## 5   400   499    43      320   0.128         0.955
## 6   500   599    15      335   0.0448        1
```

Com os dados organizados podemos construir o histograma:

```
lampadas %>%
  ggplot(aes(x = min + 50, y = freq)) +
    geom_col(
      fill = "grey",
      color = "grey30"
    ) +
```

```
  scale_x_continuous(
    breaks = seq(0, 600, 100)
  ) +
  labs(
    title = bquote(bold("Figura 3.69:")~"Tempo de vida das lâmpadas"),
    x = "Tempo de vida (h)",
    y = "Frequência observada"
  ) +
  tema
```

Figura 3.69: Tempo de vida das lâmpadas



Para calcular o quantil de ordem p, vamos seguir dois passos:

1. utilizar a tabela de frequência para identificar a classe k à qual pertence o quantil de ordem p;
2. utilizar a fomula

$$Q_p = l_k + \frac{(p \cdot n - F_{k-1}) \cdot h_k}{f_k}$$

onde:

a. l_i é o limite inferior da classe i;
b. f_i é a frequência absoluta da classe i;
c. F_i é a frequência acumulada até a classe i;
d. h_i é a amplitude da classe i;
e. n é o tamanho da amostra;

Vamos definir uma função para calcular o quantil:

```
quantil <- function(l, p, n, f_p, f_ant, h) {
  l + (((p * n) - f_ant) * (h / f_p))
}
```

Agora temos condições de calcular os quantis:

```
n <- sum(lampadas$freq)
h <- 100

tibble(
  p = c(.1, .3, .5, .7, .9),
  l = c(0, 100, 200, 300, 400),
  f_p = c(82, 71, 68, 56, 43),
  f_ant = c(0, 82, 153, 221, 277)
) %>%
  mutate(
    Q_p = quantil(l, p, n, f_p, f_ant, h)
  ) %>%
  select(p, Q_p) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.12:** Quantis para a vida útil de lâmpadas",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.13: Tabela 3.12: Quantis para a vida útil de lâmpadas

p	Q_p
0,1	40,85
0,3	126,06
0,5	221,32
0,7	324,11
0,9	456,98

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 auxilio em duas estradas num determinado dia.

Tabela 3.14: Tabela 3.13: Horários das chamadas de auxílio em duas estradas num dia específico

Estrada 1	12:07:00 AM	12:58:00 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	

a) Construa um histograma para a distribuição de frequências dos instantes de chamados em cada uma das estradas.
b) 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. Em primeiro lugar, precisamos arrumar os dados em um formato mais tangível.

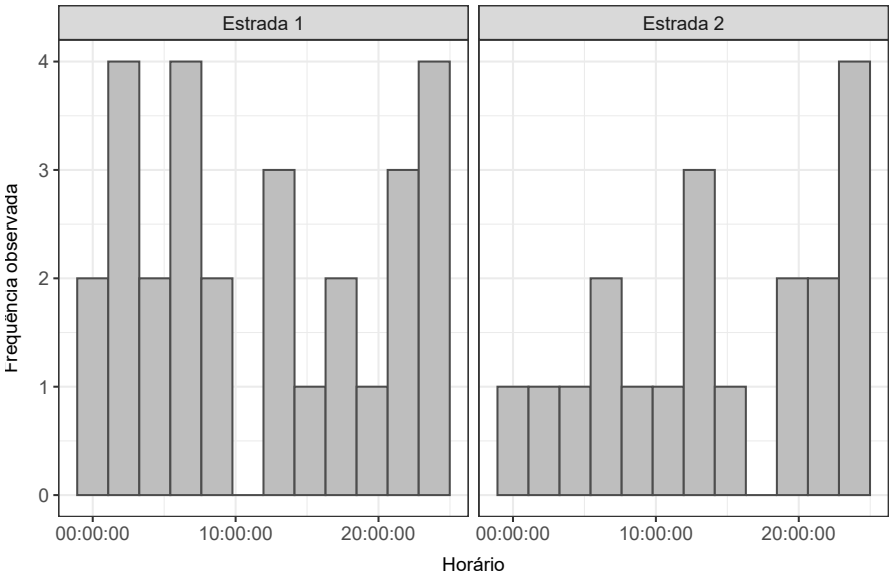
```
chamadas <- tribble(
  ~estrada, ~horario,
  "Estrada 1", "00:07:00 AM",
  "Estrada 1", "00:58:00 AM",
  "Estrada 1", "01:24:00 AM",
  "Estrada 1", "01:35:00 AM",
  "Estrada 1", "02:05:00 AM",
  "Estrada 1", "03:14:00 AM",
  "Estrada 1", "03:25:00 AM",
  "Estrada 1", "03:46:00 AM",
  "Estrada 1", "05:44:00 AM",
  "Estrada 1", "05:56:00 AM",
  "Estrada 1", "06:36:00 AM",
  "Estrada 1", "07:26:00 AM",
  "Estrada 1", "07:48:00 AM",
  "Estrada 1", "09:13:00 AM",
  "Estrada 1", "12:05:00 PM",
  "Estrada 1", "12:48:00 PM",
  "Estrada 1", "01:21:00 PM",
  "Estrada 1", "02:22:00 PM",
  "Estrada 1", "05:30:00 PM",
  "Estrada 1", "06:00:00 PM",
  "Estrada 1", "07:53:00 PM",
  "Estrada 1", "09:15:00 PM",
  "Estrada 1", "09:49:00 PM",
  "Estrada 1", "09:59:00 PM",
  "Estrada 1", "10:53:00 PM",
  "Estrada 1", "11:27:00 PM",
  "Estrada 1", "11:49:00 PM",
  "Estrada 1", "11:57:00 PM",
  "Estrada 2", "00:03:00 AM",
  "Estrada 2", "01:18:00 AM",
  "Estrada 2", "04:35:00 AM",
  "Estrada 2", "06:13:00 AM",
  "Estrada 2", "06:59:00 AM",
  "Estrada 2", "08:03:00 AM",
  "Estrada 2", "10:07:00 AM",
  "Estrada 2", "12:24:00 PM",
  "Estrada 2", "01:45:00 PM",
  "Estrada 2", "02:07:00 PM",
  "Estrada 2", "03:23:00 PM",
  "Estrada 2", "06:34:00 PM",
  "Estrada 2", "07:19:00 PM",
  "Estrada 2", "09:44:00 PM",
```

```
  "Estrada 2", "10:27:00 PM",
  "Estrada 2", "10:52:00 PM",
  "Estrada 2", "11:19:00 PM",
  "Estrada 2", "11:29:00 PM",
  "Estrada 2", "11:44:00 PM"
) %>%
  mutate(
    horario = parse_time(horario)
  )
```

Distribuição de frequência

```
chamadas %>%
  ggplot(aes(horario)) +
    geom_histogram(
      bins = 12,
      fill = "grey",
      color = "grey30"
    ) +
    facet_wrap(~ estrada) +
    labs(
      title = bquote(bold("Figura 3.70:")~"Horário das chamadas de apoio"),
      x = "Horário",
      y = "Frequência observada"
    ) +
    tema
```

Figura 3.70: Horário das chamadas de apoio



Intervalo entre as chamadas

Primeiro precisamos calcular a diferença de tempo entre as chamadas. Para isso vamos adicionar uma coluna em nosso conjunto de dados contendo o horário da chamada anterior e outra contendo a diferença.

```
chamadas$anterior = parse_time("00:00:00 AM")

n_1 <- chamadas %>%
          filter(estrada == "Estrada 1") %>%
          nrow()

n_2 <- chamadas %>%
          filter(estrada == "Estrada 2") %>%
          nrow()

for (i in 2:n_1) {
  chamadas$anterior[i] <- chamadas$horario[i-1]
}

for (i in 2:n_2) {
  chamadas$anterior[i + n_1] <- chamadas$horario[i + n_1 -1]
}
```

```
chamadas$intervalo <- as.duration(chamadas$horario - chamadas$anterior)
```

A tabela a seguir mostra o resultado desses cálculos.

```
chamadas %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.14:** Chamadas de auxílio em duas estradas",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.15: Tabela 3.14: Chamadas de auxílio em duas estradas

estrada	horario	anterior	intervalo
Estrada 1	00:07:00	00:00:00	420s (~7 minutes)
Estrada 1	00:58:00	00:07:00	3060s (~51 minutes)
Estrada 1	01:24:00	00:58:00	1560s (~26 minutes)
Estrada 1	01:35:00	01:24:00	660s (~11 minutes)
Estrada 1	02:05:00	01:35:00	1800s (~30 minutes)
Estrada 1	03:14:00	02:05:00	4140s (~1.15 hours)
Estrada 1	03:25:00	03:14:00	660s (~11 minutes)
Estrada 1	03:46:00	03:25:00	1260s (~21 minutes)
Estrada 1	05:44:00	03:46:00	7080s (~1.97 hours)
Estrada 1	05:56:00	05:44:00	720s (~12 minutes)
Estrada 1	06:36:00	05:56:00	2400s (~40 minutes)
Estrada 1	07:26:00	06:36:00	3000s (~50 minutes)
Estrada 1	07:48:00	07:26:00	1320s (~22 minutes)
Estrada 1	09:13:00	07:48:00	5100s (~1.42 hours)
Estrada 1	12:05:00	09:13:00	10320s (~2.87 hours)
Estrada 1	12:48:00	12:05:00	2580s (~43 minutes)
Estrada 1	13:21:00	12:48:00	1980s (~33 minutes)
Estrada 1	14:22:00	13:21:00	3660s (~1.02 hours)
Estrada 1	17:30:00	14:22:00	11280s (~3.13 hours)
Estrada 1	18:00:00	17:30:00	1800s (~30 minutes)
Estrada 1	19:53:00	18:00:00	6780s (~1.88 hours)
Estrada 1	21:15:00	19:53:00	4920s (~1.37 hours)
Estrada 1	21:49:00	21:15:00	2040s (~34 minutes)

estrada	horario	anterior	intervalo
Estrada 1	21:59:00	21:49:00	600s (~10 minutes)
Estrada 1	22:53:00	21:59:00	3240s (~54 minutes)
Estrada 1	23:27:00	22:53:00	2040s (~34 minutes)
Estrada 1	23:49:00	23:27:00	1320s (~22 minutes)
Estrada 1	23:57:00	23:49:00	480s (~8 minutes)
Estrada 2	00:03:00	00:00:00	180s (~3 minutes)
Estrada 2	01:18:00	00:03:00	4500s (~1.25 hours)
Estrada 2	04:35:00	01:18:00	11820s (~3.28 hours)
Estrada 2	06:13:00	04:35:00	5880s (~1.63 hours)
Estrada 2	06:59:00	06:13:00	2760s (~46 minutes)
Estrada 2	08:03:00	06:59:00	3840s (~1.07 hours)
Estrada 2	10:07:00	08:03:00	7440s (~2.07 hours)
Estrada 2	12:24:00	10:07:00	8220s (~2.28 hours)
Estrada 2	13:45:00	12:24:00	4860s (~1.35 hours)
Estrada 2	14:07:00	13:45:00	1320s (~22 minutes)
Estrada 2	15:23:00	14:07:00	4560s (~1.27 hours)
Estrada 2	18:34:00	15:23:00	11460s (~3.18 hours)
Estrada 2	19:19:00	18:34:00	2700s (~45 minutes)
Estrada 2	21:44:00	19:19:00	8700s (~2.42 hours)
Estrada 2	22:27:00	21:44:00	2580s (~43 minutes)
Estrada 2	22:52:00	22:27:00	1500s (~25 minutes)
Estrada 2	23:19:00	22:52:00	1620s (~27 minutes)
Estrada 2	23:29:00	23:19:00	600s (~10 minutes)
Estrada 2	23:44:00	23:29:00	900s (~15 minutes)

Com essas informações em mãos, podemos calcular as medidas de resumo para cada uma das estradas.

```
summ_data <- function(data) {
  x <- data$intervalo

  tibble(
     n = sum(!is.na(x), na.rm = TRUE),
    `Mínimo` = as.duration(min(x, na.rm = TRUE)),
    Q1 = as.duration(quantile(x, 0.25, na.rm = TRUE)[[1]]),
    `Mediana` = as.duration(median(x, na.rm = TRUE)),
    Q3 = as.duration(quantile(x, 0.75, na.rm = TRUE)[[1]]),
    `Máximo` = as.duration(max(x, na.rm = TRUE)),
    `Média` = as.duration(mean(x, na.rm = TRUE)),
    `Desvio Padrão` = as.duration(sd(x, na.rm = TRUE)),
    `Distância Interquartil` = as.duration(IQR(x, na.rm = TRUE))
```

```
  )
}

chamadas %>%
  group_by(estrada) %>%
  nest() %>%
  mutate(
    resumo = map(data, summ_data),
  ) %>%
  select(-data) %>%
  unnest(resumo) %>%
  t() %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.15:** Resumo do intervalo entre chamadas para as duas estradas",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.16: Tabela 3.15: Resumo do intervalo entre chamadas para as duas estradas

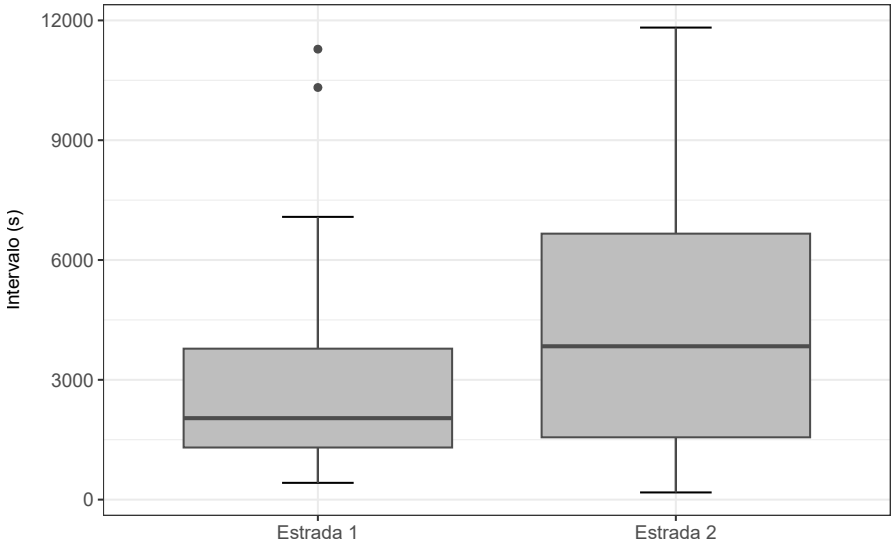
estrada	Estrada 1	Estrada 2
n	28	19
Mínimo	420s (~7 minutes)	180s (~3 minutes)
Q1	1305s (~21.75 minutes)	1560s (~26 minutes)
Mediana	2040s (~34 minutes)	3840s (~1.07 hours)
Q3	3780s (~1.05 hours)	6660s (~1.85 hours)
Máximo	11280s (~3.13 hours)	11820s (~3.28 hours)
Média	3079.28571428571s (~51.32 minutes)	4496.842105263 16s (~1.25 hours)
Desvio Padrão	2820.26256798564s (~47 minutes)	3565.35965558655s (~59.42 minutes)
Distância Interquartil	2475s (~41.25 minutes)	5100s (~1.42 hours)

O gráfico a seguir mostra a distribuição do intervalo entre chamadas:

```
chamadas %>%
  ggplot(aes(x = estrada, y = intervalo)) +
```

```
  stat_boxplot(
    geom = "errorbar",
    width = 0.2
  ) +
  geom_boxplot(
    fill = "grey",
    color = "grey30"
  ) +
  labs(
    title = bquote(bold("Figura 3.71:")~"Distribuição dos intervalos entre as chamadas"),
    x = "",
    y = "Intervalo (s)"
  ) +
  tema
```

Figura 3.71: Distribuição dos intervalos entre as chamadas



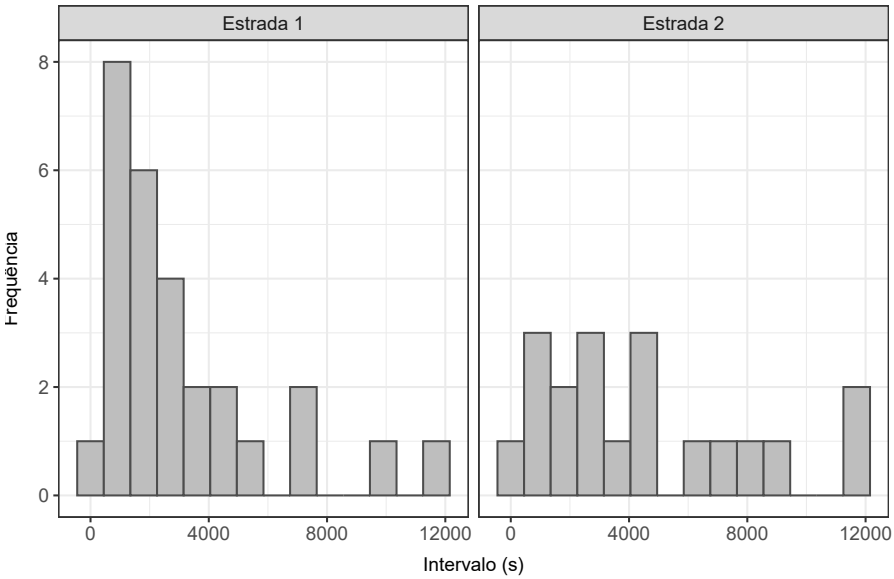
O intervalo entre as chamadas segue alguma tendência normal ou exponencial? {-}

Vamos começar plotando o histograma dos intervalos de tempo:

```
chamadas %>%
  ggplot(aes(intervalo)) +
    geom_histogram(
      binwidth = 900,
```

```
    fill = "grey",
    color = "grey30"
  ) +
  facet_wrap(~ estrada) +
  labs(
    title = bquote(bold("Figura 3.72:")~"Intervalo de tempo entre as chamadas"),
    x = "Intervalo (s)",
    y = "Frequência"
  ) +
  tema
```

Figura 3.72: Intervalo de tempo entre as chamadas

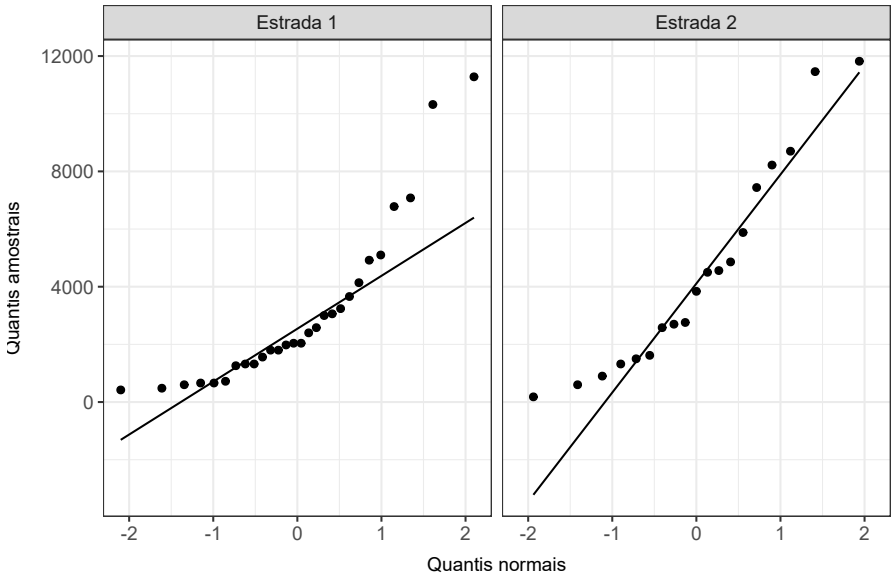


Este gráfico não nos indica nenhuma tendência, vamos então construir o gráfico QQ para cada uma das estradas, primeiro comparando com a distribuição normal:

```
chamadas %>%
  ggplot(aes(sample = intervalo)) +
    geom_qq() +
    geom_qq_line() +
    facet_wrap(~ estrada) +
    labs(
      title = bquote(bold("Figura 3.73:")~"Intervalo de tempo entre as chamadas"),
      x = "Quantis normais",
      y = "Quantis amostrais"
```

```
  ) +
  tema
```

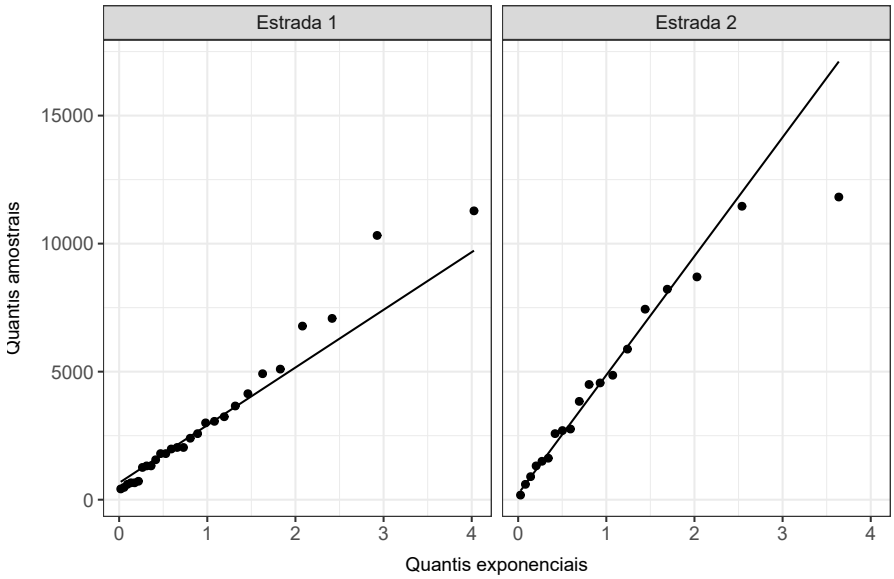
Figura 3.73: Intervalo de tempo entre as chamadas



Como os dados estão muito afastados da reta de referência, não podemos dizer que os seguem a distribuição normal. Vamos tentar agora compará-los a uma distribuição exponencial.

```
chamadas %>%
  ggplot(aes(sample = intervalo)) +
    geom_qq(
      distribution = stats::qexp
    ) +
    geom_qq_line(
      distribution = stats::qexp
    ) +
    facet_wrap(~ estrada) +
    labs(
      title = bquote(bold("Figura 3.74:")~"Intervalo de tempo entre as chamadas"),
      x = "Quantis exponenciais",
      y = "Quantis amostrais"
    ) +
    tema
```

Figura 3.74: Intervalo de tempo entre as chamadas



Notamos que os dados estão muito mais próximos da distribuição exponencial, exceto por alguns pontos extremos.

Exercício 3.16

As notas finais de um curso de Estatistica 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.

a) Calcule a mediana, os quartis e a média.
b) 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 variáncia das notas desses dois grupos.

Solução. Vamos organizar as notas em um data frame e incluir o status (Aprovado ou Reprovado) aos estudantes.

```
estatistica <- tibble(
                  nota = c(7, 5, 4, 5, 6, 3, 8, 4, 5, 4, 6, 4, 5, 6, 4, 6, 6, 3, 8, 4, 5, 4, 5, 5, 6)
              ) %>%
              mutate(
                status = ifelse(nota < 5, "Reprovado", "Aprovado")
              )
```

```
head(estatistica)
```

```
## # A tibble: 6 x 2
##    nota status
##   <dbl> <chr>
## 1     7 Aprovado
## 2     5 Aprovado
## 3     4 Reprovado
## 4     5 Aprovado
## 5     6 Aprovado
## 6     3 Reprovado
```

Agora podemos calcular as medidas de resumo:

```
estatistica$nota %>%
  summ() %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.16:** Medidas de resumo para as notas em estatística",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.17: Tabela 3.16: Medidas de resumo para as notas em estatística

	x
n	25,00
Mínimo	3,00
Q1	4,00
Mediana	5,00
Q3	6,00
Máximo	8,00
Média	5,12
Desvio Padrão	1,33
Distância Interquartil	2,00

Vamos agora comparar os grupos:

```
apr <- estatistica %>% filter(status == "Aprovado")
rep <- estatistica %>% filter(status == "Reprovado")
```

A tabela a seguir mostra um resumo dos aprovados:

```
apr$nota %>%
  summ() %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.17:** Medidas de resumo para as notas dos aprovados em estatística",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 3.18: Tabela 3.17: Medidas de resumo para as notas dos aprovados em estatística

	x
n	16,00
Mínimo	5,00
Q1	5,00
Mediana	6,00
Q3	6,00
Máximo	8,00
Média	5,88
Desvio Padrão	1,02
Distância Interquartil	1,00

E a próxima tabela exibe um resumo para os reprovados:

```
rep$nota %>%
  summ() %>%
  kable(
    format = "pipe",
    caption = "**Tabela 3.18:** Medidas de resumo para as notas dos reprovados em estatística",
    label = NA,
    digits = 2,
```

```
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

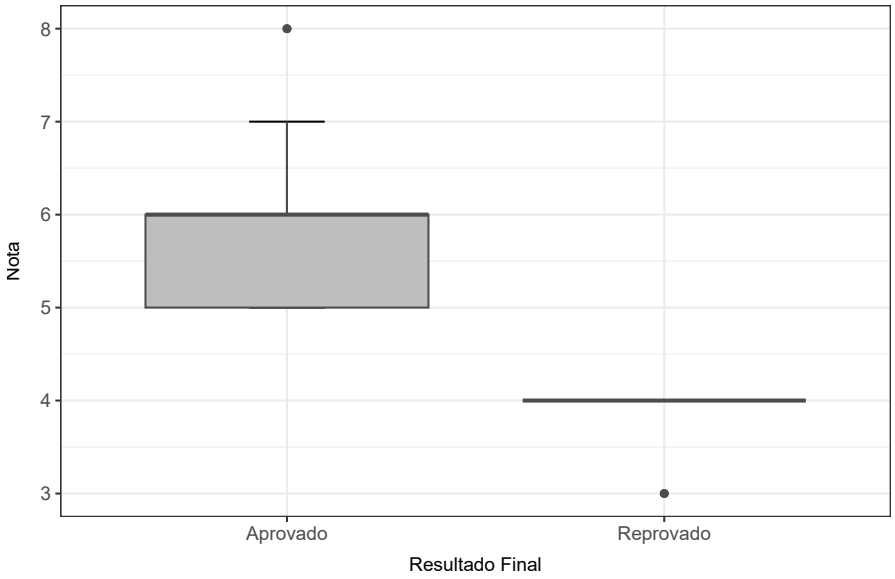
Tabela 3.19: Tabela 3.18: Medidas de resumo para as notas dos reprovados em estatística

	x
n	9,00
Mínimo	3,00
Q1	4,00
Mediana	4,00
Q3	4,00
Máximo	4,00
Média	3,78
Desvio Padrão	0,44
Distância Interquartil	0,00

Avaliando o desvio padrão, e o intervalo interquartil, percebemos que as notas dos reprovados são muito mais uniformes, isto é, apresentam menor variação. O mesmo pode ser constatado no gráfico a seguir.

```
estatistica %>%
  ggplot(aes(x = status, y = nota)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = bquote(bold("Figura 3.75:")~"Distribuição das notas em Estatística"),
      x = "Resultado Final",
      y = "Nota"
    ) +
    tema
```

Figura 3.75: Distribuição das notas em Estatística



Exercício 3.17

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

Tabela 3.20: Tabela 3.19: Pulsação de estudantes durante atividade física

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:

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

Solução.

Item a

A tabela nos informa que o primeiro quartil entre o grupo que praticou atividade intensa é 70, logo, espera-se que 25% dos estudantes neste grupo tiveram pulsação inferior a 70 e não 5% como alegado no enunciado. Entre os estudantes que praticaram atividade fraca, como a mediana é 70, espera-se que pelo menos a metade tenha pulsação menor do que 70 (note contudo, que podem haver mais o que 50% dos estudantes!).

Conclusão: o item é FALSO.

Item b

Considerando que o primeiro quartil vale 70 para o grupo de estudantes com atividade intensa e 63 para o grupo com atividade fraca, sabemos que pelo menos 25% dois estudantes de cada grupo estão abaixo desses valores.

Conclusão: a afirmação é FALSA.

item c

Considerando a o intervalo de confiança para a média como $[\bar{x} - s; \bar{x} + s]$, temos que a verdadeira média para o grupo com atividade intensa está entre 69,1 e 90,1 e, para o segundo grupo, entre 63,5 e 82,7. Como os intervalos não são disjuntos, não é possível afirmar que há diferença entre as médias.

Conclusão: a afirmativa é VERDADEIRA.

item d

Como a mediana é 82, só podemos concluir que pelo menos metade dos estudantes tem pulsação maior do que ou igual a 82.

Conclusão: o item é FALSO.

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. Analisando o gráfico, a correspondencia mais adequada seria A e D, que não está listada acima.

Exercício 3.19

Os histogramas apresentados na Figura 3.35 mostram a distribuição das temperaturas (°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. Vamos iniciar a análise considerando o ponto médio e a frequência das classes:

```
r1 <- c(10,10,10,10,10,10,12,12,12,12,14,16,16,16,16,18,18,18,18,18,18)
r2 <- c(10,10,10,10,12,12,12,12,14,14,14,14,14,16,16,16,16,18,18,18,18)

mean(r1)
```

```
## [1] 14
```

```
sd(r1)
```

```
## [1] 3.34664
```

```
mean(r2)
```

```
## [1] 14
```

```
sd(r2)
```

```
## [1] 2.828427
```

Item a

Temos a mesma média (14ºC), porém desvios padrão diferentes (3,35 para R1 e 2,83 para R2).

Conclusão: A afirmação é FALSA.

Item b

Podemos calcular o desvio padrão considerando o ponto médio e as frequências das classes.

Conclusão: A afirmativa é falsa.

Item c

Conforme calculamos acima, as médias são iguais.

Conclusão: A afirmativa é falsa.

Item d

Conforme calculamos acima, as medias são iguais, mas os desvios são diferentes.

Conclusão: A afirmação é VERDADEIRA.

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.

Item a

Como o terceiro quartil é maior do que ou igual a 75% dos valores, a probabilidade de ganhar menos do que 5000 é de, pelo menos 75%. Logo é mais provável que o salário não ultrapasse este valor.

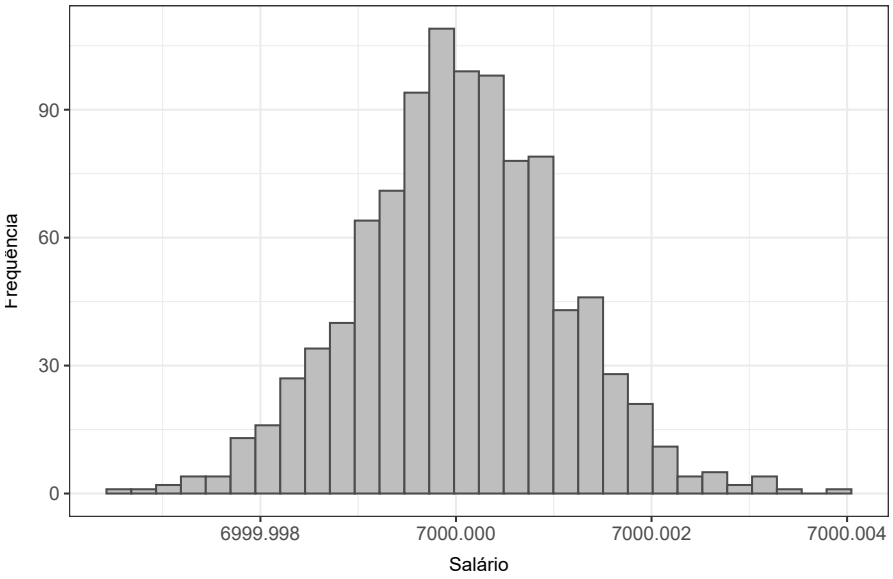
Item b

Para melhor exemplificar, vamos simular uma distribuição com média 7000 e desvio padrão igual a 1/1000.

```
salario <- tibble(
  x = rnorm(1000, mean = 7000, sd = 1/1000)
)

salario %>%
  ggplot(aes(x = x)) +
    geom_histogram(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = bquote(bold("Figura 3.76:")~"Distribuição dos salários na empresa B"),
      x = "Salário",
      y = "Frequência"
    ) +
    tema
```

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

Figura 3.76: Distribuição dos salários na empresa B



Como a dispersão é muito pequena, os valores estão super concentrados ao redor da média, neste caso, a probabilidade de o salário ser menor do que 5000 é quase nula, tornando a emprea B mais atrativa.

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.

A distância interquartis é 5

A afirmação é falsa, pois a distância interquartil é dada por $d_Q = Q_3 - Q_1$, logo, para o nosso exemplo, $d_Q = 20 - 10 = 10$.

O valor 32 seria considerado *outlier* segundo o critério utilizado na construção do *boxplot*

Os outliers são aqueles que distam mais do que $1,5 \cdot d_Q$ dos quartis 1 e 3. Dessa forma, seria considerados *outliers* os valores abaixo de -5 ($Q_1 - 1,5 \cdot d_q$) ou acima de 35 ($Q_3 + 1,5 \cdot d_Q$), tornando a afirmação é falsa.

A mediana ficaria alterada de 2 unidades se um ponto com valor acima do terceiro quartil fosse substituído por outro 2 vezes maior

Como a mediana divide o conjunto em partes de igual tamanho, substituir valores acima da mediana por outros maiores não muda o seu valor. A afirmação é, então, uma falsidade.

O valor mínimo é maior do que zero

Não há como afirmar!

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.

Os medicamentos são equivalentes?

Como os desvios padrão são diferentes, os medicamentos NÃO são equivalentes.

A probabilidade de cura em até 40s é igual para ambos?

Considerando que o tempo de cura se distribui normalmente, vamos utilizar a função `pnorm()` para calcular a probabilidade de cura para ambos os medicamentos:

```
(p_A <- pnorm(40, 60, 10))
```

```
## [1] 0.02275013
```

```
(p_B <- pnorm(40, 60, 30))
```

```
## [1] 0.2524925
```

Como `p_A` < `p_B`, a probabilidade de a dor de cabeça ser curada em até 40 segundos é maior para o medicamento B.

Obs: A suposição de normalidade foi apenas para facilitar os calculos, mas o resultado se repetiria para qualquer distribuição simetrica centrada e com pico na média.

A probabilidade de cura em até 60s é maior para o medicamento B?

Como, para ambos os medicamentos, se supôs a simetria e centralidade na média, as probabilidades de cura até a média são iguais em ambos os medicamentos e correspondem à 50%.

Exercício 3.23

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

Tabela 3.21: Tabela 3.20: Número de dependentes poe empregado

Dependentes	Frequência
1	40
2	50
3	30
4	20
5	10
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. Para este caso, a média é dada por

$$\frac{\sum x_i f_i}{\sum f_i}$$

onde:

- x_i representa o número de dependentes na linha i;
- f_i representa o número de empregados que possuem x_i dependentes;

Dessa forma, temos:

$$\bar{x} = \frac{(1 \cdot 40) + (2 \cdot 50) + (3 \cdot 30) + (4 \cdot 20) + (5 \cdot 10)}{40 + 50 + 30 + 20 + 10} = \frac{360}{150} = 2,4$$

Para a mediana e a moda, vamos calcular as frequências relativa e acumulada na tabela.

Tabela 3.22: Tabela 3.21: Número de dependentes por empregado

Dependentes	Frequência	Frequência relativa (%)	Frequência acumulada (%)
1	40	26,67	26,67
2	50	33,33	60,00
3	30	20,00	80,00
4	20	13,33	93,33
5	10	6,67	100,00
Total	150	100,00	100,00

A mediana corresponderá ao valor x_i na linha i que contém frequência acumulada maior do que ou igual a 50% e a moda será o valor x_j da linha j que contém a maior frequência ou frequência relativa.

Nesse caso, temos:

- mediana igual a 2 filhos;
- moda igual a 2 filhos;
- média igual a 2,4.

E nenhuma das alternativas está correta.

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. Somando as frequências relativas correspondentes, chegamos ao valor de 73,33%, indicado pela alternativa (d).

Exercício 3.25

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

Tabela 3.23: Tabela 3.22: Número de vasos alimentando um tumor

Número de vasos	Frequência
$0 \vdash 5$	8 (12%)
$5 \vdash 10$	23 (35%)
$10 \vdash 15$	12 (18%)
$15 \vdash 20$	9 (14%)
$20 \vdash 25$	8 (12%)
$25 \vdash 30$	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. Para facilitar a visualização, vamos separar a frequência relativa e adicionar uma nova coluna com a frequência acumulada.

Tabela 3.24: Tabela 3.23: Número de vasos alimentando um tumor

Número de vasos	Frequência	Frequência relativa (%)	Frequência acumulada(%)
$0 \vdash 5$	8	12	12
$5 \vdash 10$	23	35	47
$10 \vdash 15$	12	18	65
$15 \vdash 20$	9	14	79
$20 \vdash 25$	8	12	91
$25 \vdash 30$	6	9	100
Total	66	100	100

O primeiro quartil é 25%?

Para calcular o valor do primeiro quartil, vamos usar a função `quantil()` que criamos anteriormente e chegamos ao valor de 8.6956522, mostrando que a afirmação é FALSA.

A mediana está entre 10 e 15?

Basta encontrarmos a classe que contém frequência acumulada maior ou igual a 50%. No nosso caso, a classe é a que vai de 10 a 15 vasos. A alternativa é VERDADEIRA.

O percentil de ordem 10% é 10?

Como a primeira classe, que vai de 0 a 5 vasos, acumula 12% das observações, concluímos que o primeiro decil está entre 0 e 5, tornando a afirmação FALSA.

A distância interquartis é 50?

Como nossa variável está distribuida entre 0 e 30, não há como a distância interquartil ser maior do que 30. A afirmação é FALSA.

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 variancia 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.

Não é possível estimar nem a média nem a variância com esses dados?

Conforme vimos no Exercício 3.23, a média pode ser estimada. No item a seguir, veremos que também é possível estimar a variância, logo a afirmação FALSA.

A variancia é menor que 30?

Podemos estimar a variância utilizando o ponto médio das classes através da seguinte equação:

$$s^2 = \frac{\sum \left[f_i \cdot (x_i - \bar{x})^2 \right]}{n-1}$$

Em nosso exemplo, temos a $s^2 \approx 57.60$, o que torna a afirmação FALSA.

A média estimada é 12,8?

Usando a formula acima, temos que $\bar{x} \approx 12.80$, logo a afirmação é VERDADEIRA.

Em apenas 35% dos casos, o número de vasos é maior que 10?

Basta somarmos as frequências relativas das classes correspondentes para notar que em 53% dos casos, o número de vasos é igual ou superior a 10. A afirmação é, portanto, FALSA.

Exercício 3.27

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

Tabela 3.25: Tabela 3.24: 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.

Não é possível estimar o nível médio de colesterol populacional só com esses dados

FALSO, porque a média amostral é uma estimativa da média populacional, além disso os desvios padrão nos ajudam a calcular um intervalo de confiança para a média populacional.

Se os dois estudos foram realizados com amostras da mesma população não deveria haver diferença entre os desvios padrões amostrais

FALSO, porque como o desvio padrão é calculado a partir da amostra, pode ocorrer diferença entre amostras diferentes.

Com os dados do estudo B, o colesterol médio populacional pode ser estimado com mais precisão do que com os dados do estudo A

VERDADEIRO?

Ambos os estudos sugerem que a distribuição do colesterol na população é simétrica

FALSO, pois só com esses dados não há como afirmar.

Exercício 3.28

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

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

Solução.

Calculando a média de W_i

$$\begin{aligned}
\overline{W} &= \frac{1}{n} \cdot \sum_{i=1}^{n} W_i \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i + k) \\
&= \frac{1}{n} \cdot \left[\left(\sum_{i=1}^{n} k \right) + \left(\sum_{i=1}^{n} X_i \right) \right] \\
&= \frac{1}{n} \cdot \left[(n \cdot k) + \left(\sum_{i=1}^{n} X_i \right) \right] \\
&= \left(\frac{1}{n} \cdot n \cdot k \right) + \left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i \right) \\
&= k + \overline{X}
\end{aligned}$$

Calculando o desvio padrão de W_i

$$\begin{aligned}
s_W^2 &= \frac{1}{n} \cdot \sum_{i=1}^{n} (W_i - \overline{W})^2 \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left[(X_i + k) - \left(k + \overline{X} \right) \right]^2 \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(X_i + k - k - \overline{X} \right)^2 \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(X_i + \overline{X} \right)^2 \\
&= s_X^2
\end{aligned}$$

Logo $s_W = s_X$, isto é, os desvios padrão são iguais.

Calculando a média de V_i

$$\begin{aligned}
\overline{V} &= \frac{1}{n} \cdot \sum_{i=1}^{n} V_i \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} (k \cdot X_i) \\
&= \frac{1}{n} \cdot k \cdot \sum_{i=1}^{n} X_i \\
&= \frac{1}{n} \cdot \left[(n \cdot k) + \left(\sum_{i=1}^{n} X_i \right) \right] \\
&= k \cdot \overline{X}
\end{aligned}$$

Calculando o desvio padrão de V_i

$$\begin{aligned}
s_V^2 &= \frac{1}{n} \cdot \sum_{i=1}^{n} (V_i - \overline{V})^2 \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left[(k \cdot X_i) - \left(k \cdot \overline{X} \right) \right]^2 \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left[k \cdot \left(X_i - \overline{X} \right) \right]^2 \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left[k^2 \cdot \left(X_i - \overline{X} \right)^2 \right] \\
&= \frac{1}{n} \cdot k^2 \cdot \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 \\
&= k^2 \cdot s_X^2 \\
&= (k \cdot s_X)^2
\end{aligned}$$

Logo $s_V = k \cdot s_X$.

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

Solução. Para a média da variável z, temos:

$$\begin{aligned}
\overline{z} &= \frac{1}{n} \cdot \sum_{i=1}^{n} z_i \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \\
&= \frac{1}{n} \cdot \frac{1}{s_x} \cdot \sum_{i=1}^{n} (x_i - \overline{x}) \\
&= \frac{1}{n} \cdot \frac{1}{s_x} \cdot \left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} \right] \\
&= \frac{1}{n} \cdot \frac{1}{s_x} \cdot [n\overline{x} - n\overline{x}] \\
&= \frac{1}{n} \cdot \frac{1}{s_x} \cdot 0 \\
&= 0
\end{aligned}$$

Para o desvio padrão de z, temos:

$$\begin{aligned}
s_z^2 &= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (z_i - \overline{z}) \\
&= \frac{1}{n-1} \cdot \sum_{i=1}^{n} z_i \\
&= \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right)^2 \\
&= \frac{1}{n-1} \cdot \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{s_x^2} \\
&= \frac{1}{n-1} \cdot \frac{1}{s_x^2} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= \frac{1}{s_x^2} \cdot \left[\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \right] \\
&= \frac{1}{s_x^2} \cdot s_x^2 \\
&= 1
\end{aligned}$$

Segue que $s_z = 1$.

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 $X_1, \cdots, X_n$ 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 tamanhon $n = 4$ 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 quiquadrado com 3 graus de liberdade.

Solução. x

4

Análise de dados de duas variáveis

4.1 Introdução

De maneira geral, dizemos que existe uma associação entre duas variáveis se o conhecimento do valor de uma delas nos dá alguma característica da distribuição (de frequência) da outra. [Morettin and Singer, 2022, p. 79]

4.2 Duas variáveis qualitativas

Dnotando os valores observados por o_i e os esperados por e_i, $i = 1, 2, 3, 4$, podemos calcular os *residuos* $r_i = o_i - e_i$ e verificar que $\sum_i r_i = 0$. Uma medida da discrepância entre os valores observados e aqueles esperados sob a hipótese H é chamada estatística ou **qui-quadrado** de Pearson,

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

[Morettin and Singer, 2022, p. 82].

A própria estatística de Pearson poderia servir como medida da intensidade da associação mas seu valor valor aumenta com o tamanho da amostra; uma alternativa para corrigir ese problema é o **coeficiente de contingência de Pearson**, dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

[Morettin and Singer, 2022, p. 83].

Uma modificação de C é o ** ceficiente de Tschuprov**,

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(c-1)}}},$$

que atinge o valor máximo igual a 1 quando $r = c$ [Morettin and Singer, 2022, p. 83].

Uma estimativa do índice denominado κ de Cohen (1960), construido com esse propósito [avaliar a magnitude da concordância entre duas classificações] é

$$\widehat{\kappa} = \frac{\sum_{i=1}^{3} p_{ii} - \sum_{i=1}^{3} p_{i+}p_{+i}}{1 - \sum_{i=1}^{3} p_{i+}p_{+i}}$$

Nessa expressão, $p_{ij} representa frequência relativa associada à casela correspondente à linha i e coluna j da tabela [de contingência] e p_{i+} e p_{+j} representam a soma das frequências relativas asociadas à linha i e coluna j, respectivamente [Morettin and Singer, 2022, p. 85].

Risco atribuível: $d = \pi_1 - \pi_0$, que coresponde a diferença entre as probabilidades (ou riscos) de ocorrência do evento de interese para expostos e não expostos ao fator de risco [Morettin and Singer, 2022, p. 86].

Risco relativo: $r = \pi_1/\pi_0$, que corresponde ao quociente entre as probabilidades de ocorrência do evento de interesse para expostos e não expostos ao fator de risco [Morettin and Singer, 2022, p. 87].

Razão de chances (*odds ratio*): $\omega = [\pi_1/(1-\pi_1)]/[\pi_0/(1-\pi_0)]$, que corresponde ao quociente entre as chances de ocorrência do evento de interesse para expostos e não expostos ao fator de risco [Morettin and Singer, 2022, p. 87].

Acurácia: corresponde à probabilidade de resultados corretos [Morettin and Singer, 2022, p. 89].

Sensibilidade e especificidade são características do teste, mas tanto o valor preditivo positivo quanto o valor preditivo negativo dependem da **prevalência** (porcentagem de indivíduos doentes na população) da doença [Morettin and Singer, 2022, p. 89].

4.3 Duas variáveis quantitativas

Dado um conjunto de n pares (x_i, y_i), a associação (linear) entre as variáveis quantitativas X e Y pode ser quantificada por meio do **coeficiente de correlação (linear)** de Pearson, definifo por

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2\right]^{1/2}}$$

[Morettin and Singer, 2022, p. 91].

[...] uma medida de associação mais robusta é o coeficiente de correlação de Apearman cuja expressão é similar à (4.4) [coef. de Pearson] com valores de X e Y substituídos pelos respectivos **postos**. [...]

$$r_p = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\left[\sum_{i=1}^{n}(R_i - \overline{R})^2 \cdot \sum_{i=1}^{n}(S_i - \overline{S})^2\right]^{1/2}}$$

[Morettin and Singer, 2022, p. 92].

[...] o coeficiente de correlação de Spearman é mais apropriado para avaliação de associações não lineares, desde que sejam **monotônicas** [...] [Morettin and Singer, 2022, p. 93].

[...] dados longitudinais, *i.e.*, aqueles qem que a mesma variável é observada em cada unidade amostral mais do que uma vez ao longo do tempo [...] [Morettin and Singer, 2022, p. 95].

Gráfico de médias/diferenças ou gráfico de Bland-Altman "consiste num gráfico das diferenças entre duas observações pareadas $(X_{2i} - X_{1i})$ em função das médias correspondentes $[(X_{1i} + X_{2i})/2]$, $i = 1, \cdots, n$" [Morettin and Singer, 2022, p. 97-98].

4.4 Uma variável qualitativa e outra quantitativa

Outro gráfico útil para avaliar a associação entre a variável quantitativa [...] e a variável qualitativa [...] é o **gráfico de perfis médios**. Nesse gráfico cartesiano as médias (e barras representando desvios padrões, erros padrões ou intervalos de confiança [...]) da variável quantitativa são representadas no eixo das ordenadas e os níveis da variável qualitativa, no eixo das abscissas [Morettin and Singer, 2022, p. 101].

Então podemos definir o grau de associação entre duas variáveis como o ganho relativo na variância obtido pela introdução da variável qualitativa. Explicitamente:

$$R^2 = \frac{Var(S) - \overline{Var(S)}}{Var(S)} = 1 - \frac{\overline{Var(S)}}{Var(S)}$$

Além disso, pode-se mostrar que $0 \leq R^2 \leq 1$ [Morettin and Singer, 2022, p. 106].

R^2 indica o percentual da variação de uma variável dependente que pode ser explicada pela variável independente [Morettin and Singer, 2022, p. 106].

4.5 Notas de capítulo

Embora não se possa calcular o risco relativo de doença em estudos retrospectivos, a razão de chances obtida por meio desse tipo de estudo é igual àquela que seria obtida por intermédio de um estudo

prospectivo, que em muitas situações práticas não pode ser realizado devido ao custo [Morettin and Singer, 2022, p. 107].

4.6 Exercícios

Exercício 4.1

Considere o conjunto de dados disponível no arquivo `empresa.xls`. Compare as distribuições de frequências das variáveis `Estado civil`, `Grau de Instrução` e `Salário` para indivíduos com diferentes procedências.

Solução. Vamos começar carregando os dados.

```
empresa <- readxl::read_xls(paste0(data_dir, "empresa.xls"), skip = 1)
colnames(empresa) <- c("id", "estado", "instrucao", "filhos", "salario", "anos", "meses", "regiao")

empresa <- empresa %>%
              mutate(
                estado = parse_factor(estado, levels = c("solteiro", "casado", "divorciado", "viúdo"), ordered =
                instrucao = parse_factor(instrucao, ordered = TRUE, levels = c("ensino fundamental", "ensino méd
                regiao = parse_factor(regiao, ordered = FALSE, levels = c("interior", "capital", "outra")),
                idade = anos + (meses / 12)
              )

write_csv(empresa, paste0(data_dir, "empresa.csv"))
```

Com os dados carregados e arrumados, podemos partir para as análises.

Estado Civil *versus* Região de Procedência

Por se tratar de duas variáveis categóricas, vamos analisá-las a partir da tabela de contingência, à estatística qui-quadrado de Pearson e ao coeficiente de contingência de Pearson.

```
empresa %>%
  group_by(estado, regiao) %>%
```

```
  count() %>%
  bind_rows(
    group_by(., estado) %>%
      summarise(n = sum(n)) %>%
      mutate(regiao = "Total")
  ) %>%
  bind_rows(
    group_by(., regiao) %>%
      summarise(n = sum(n)) %>%
      mutate(estado = "Total")
  ) %>%
  spread(regiao, n) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.1:** Distribuição conjunta das variáveis Estado Civil e Região de Procedência",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.1: Tabela 4.1: Distribuição conjunta das variáveis Estado Civil e Região de Procedência

estado	capital	interior	outra	Total
casado	7	8	5	20
solteiro	4	4	8	16
Total	11	12	13	36

Utilizaremos a função `chisq.test()` para calcular nossas estatísticas de interesse:

```
qui_quadrado <- chisq.test(empresa$estado, empresa$regiao)$statistic[[1]]
```

```
## Warning in chisq.test(empresa$estado, empresa$regiao): Aproximação do
## qui-quadrado pode estar incorreta
```

```
n <- nrow(empresa)
c <- sqrt(qui_quadrado / (qui_quadrado + n))
```

Após os cálculos, constatamos que o valor de χ^2 é 2.4293706 e o coeficiente C de contingência de Pearson é de 0.2514289, indicando uma baixa correlação entre as variáveis.

Grau de Instrução *versus* Região de Procedência

Novamente usaremos a tabela de contingência, a estatística qui-quadrado de Pearson e o coeficiente de contingência de Pearson.

```
empresa %>%
  group_by(instrucao, regiao) %>%
  count() %>%
  bind_rows(
    group_by(., instrucao) %>%
      summarise(n = sum(n)) %>%
      mutate(regiao = "Total")
  ) %>%
  bind_rows(
    group_by(., regiao) %>%
      summarise(n = sum(n)) %>%
      mutate(instrucao = "Total")
  ) %>%
  spread(regiao, n) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.2:** Distribuição conjunta das variáveis Grau de Instrução e Região de Procedência",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.2: Tabela 4.2: Distribuição conjunta das variáveis Grau de Instrução e Região de Procedência

instrucao	capital	interior	outra	Total
ensino fundamental	4	3	5	12
ensino médio	5	7	6	18
superior	2	2	2	6
Total	11	12	13	36

Utilizaremos a função `chisq.test()` para calcular nossas estatísticas de interesse:

```
qui_quadrado <- chisq.test(empresa$instrucao, empresa$regiao)$statistic[[1]]
```

```
## Warning in chisq.test(empresa$instrucao, empresa$regiao): Aproximação do
## qui-quadrado pode estar incorreta
```

```
n <- nrow(empresa)
c <- sqrt(qui_quadrado / (qui_quadrado + n))
```

Após os cálculos, constatamos que o valor de χ^2 é 0.6614219 e o coeficiente C de contingência de Pearson é de 0.1343181, indicando uma baixa correlação entre as variáveis.

Salário *versus* Região de Procedência

Agora temos em mãos uma variável qualitativa e outra quantitativa. Utilizaremos uma tabela de resumo e gráfico de *boxplots* para avaliar a relação, além de calcularmos a estatística R^2.

A Tabela 4.3 a seguir apresenta as medidas de resumo para o salário dos funcionários, conforme a região de procedência do funcionário. A Figura 4.1 apresenta uma comparação gráfica dos salários utilizando *boxplots*.

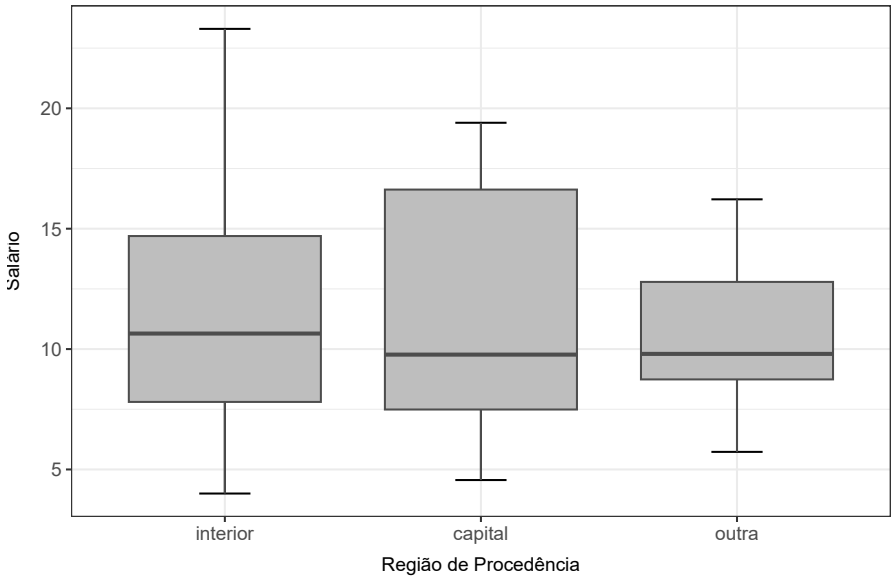
```
empresa %>%
  statds_summarise(salario, regiao) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.3:** Medidas de resumo para a variável `Salário`, conforme `Região de procedência`",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.3: Tabela 4.3: Medidas de resumo para a variável `Salário`, conforme `Região de procedência`

regiao	n	Média	Variância	Desvio Padrão	Min.	Q1	Mediana	Q3	Máx.	IQR
interior	12	11,55	28,05	5,30	4,00	7,80	10,64	14,70	23,30	6,89
capital	11	11,46	29,99	5,48	4,56	7,49	9,77	16,62	19,40	9,13
outra	13	10,45	9,89	3,15	5,73	8,74	9,80	12,79	16,22	4,05

```
empresa %>%
  ggplot(aes(regiao, salario)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = bquote(bold("Figura 4.1:")~"Distribuição dos salários conforme região de procedência"),
      x = "Região de Procedência",
      y = "Salário"
    ) +
    tema
```

Figura 4.1: Distribuição dos salários conforme região de procedência



Utilizamos a função `var()` da base do R para calcular a variância total da variável salário $Var(salrio) = 21.04477$ e os dados da Tabela 4.1 para calcular a média de variação da variável por grupos $\overline{Var(salrio)} = 22.08694$ e chegamos a $R^2 = -0.0495217$. Como os valores estão inconsistentes, será necessário revisar este item.

Exercício 4.2

Considere o conjunto de dados disponível no arquivo `regioes.xls`. Avalie a relação entre variáveis `Região` e `Densidade populacional`.

Solução. Como de costume, iniciaremos o exercício com o carregamento e ajuste dos dados.

```
regioes <- statds_read("regioes.xls", "xls")
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
```

```
colnames(regioes) <- c("regiao", "estado", "superficie", "populacao", "densidade")

regioes <- regioes %>%
              filter(estado != "Subtot") %>%
              filter(estado != "Total") %>%
              fill(regiao) %>%
              mutate(
                regiao = parse_factor(regiao),
                estado = parse_factor(estado)
              )

statds_write(regioes, "regioes.csv")
```

Como temos uma variável qualitativa e outra quantitativa, vamos exibir uma tabela de resumos, um gráfico de *boxplots* e calcular o coeficiente de determinação R^2.

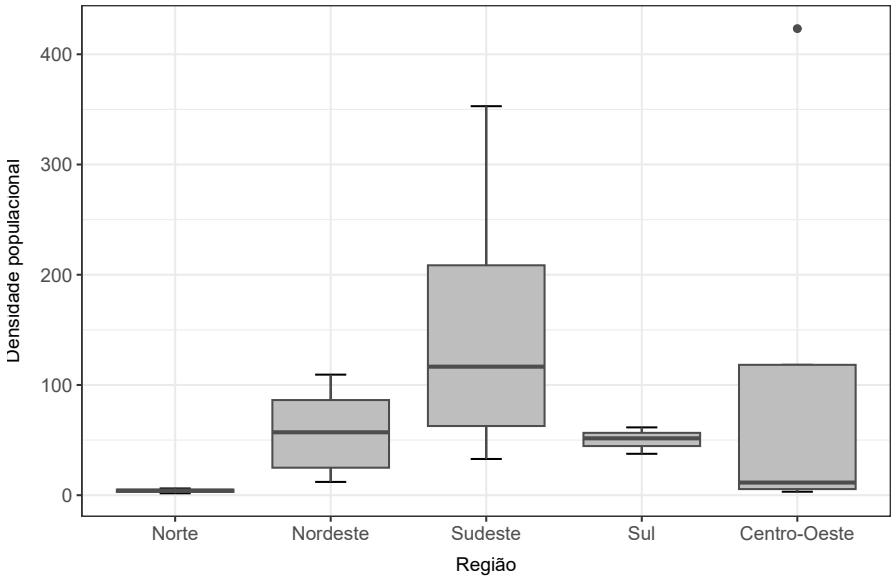
```
regioes %>%
  statds_summarise(densidade, regiao) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.4:** Medidas de resumo para a variável `Densidade Populacional`, conforme `Região`",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.4: Tabela 4.4: Medidas de resumo para a variável `Densidade Populacional`, conforme `Região`

regiao	n	Média	Variância	Desvio Padrão	Min.	Q1	Mediana	Q3	Máx.	IQR
Norte	7	4,03	2,71	1,64	1,76	3,02	4,11	5,07	6,12	2,05
Nordeste	9	57,36	1148,98	33,90	12,06	24,94	57,08	86,32	109,38	61,39
Sudeste	4	154,74	20293,27	142,45	32,86	62,77	116,60	208,57	352,90	145,80
Sul	3	50,23	145,04	12,04	37,56	44,58	51,60	56,56	61,53	11,98
Centro-Oeste	4	112,35	43003,52	207,37	3,16	5,55	11,47	118,28	423,29	112,73

```
regioes %>%
  ggplot(aes(regiao, densidade)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
    labs(
      title = bquote(bold("Figura 4.2:")~"Densidade populacional conforme região"),
      x = "Região",
      y = "Densidade populacional"
    ) +
    tema
```

Figura 4.2: Densidade populacional conforme região



Pendente calcular o R^2 e interpretar os resultados.

Exercício 4.3

Considere o conjunto de dados disponível no arquivo `salarios.xls`.

a. Compare as distribuições das variáveis `Salário de professor secundário` e `Salário de administrador` por de um gráfico QQ e interprete os resultados.
b. Calcule o coeficiente de correlação de Pearson e o coeficiente de correlação robusto (4.15) com $\alpha = 0,10$ entre essas duas variáveis.

Solução. Carregando os dados:

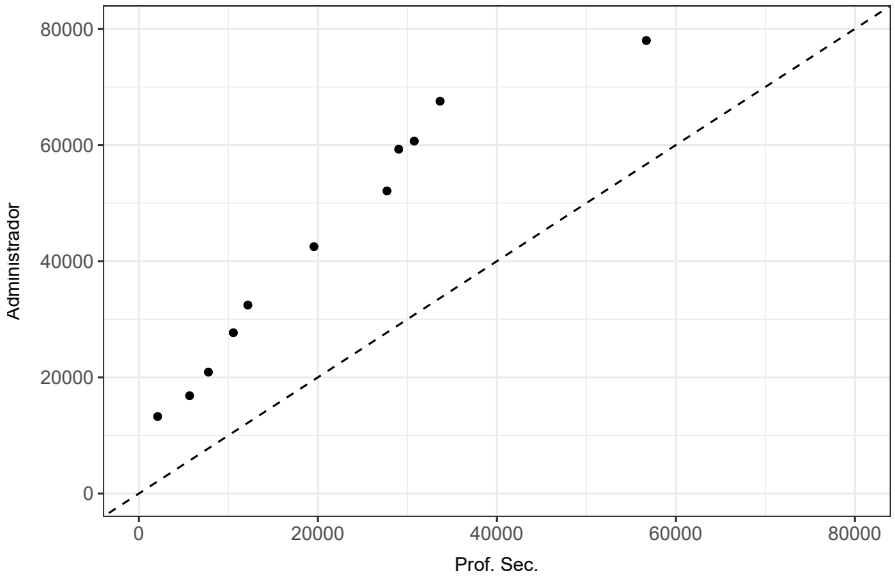
```
salarios <- statds_read("salarios.xls", "xls", skip = 4)
colnames(salarios) <- c("cidade", "prof_sec", "mecanico","administrador","eng_eletr")
```

Comparando os salários a partir de um gráfico QQ

Como a função `geom_qq()` trabalha apenas com uma variável (comparando os seus quantis com os quantis teóricos), precisaremos construir o gráfico manualmente.

```
quantiles <- seq(0, 1, 0.1)
quantiles_prof_sec <- quantile(salarios$prof_sec, quantiles)
quantiles_administrador <- quantile(salarios$administrador, quantiles)

ggplot(
  mapping = aes(
              x = quantiles_prof_sec,
              y = quantiles_administrador,
              slope = 1,
              intercept = 0
            )
) +
  geom_point() +
  geom_abline(linetype = 2) +
  labs(
    title = bquote(bold("Figura 4.3:")~"Quantis entre o salário de professores secundaristas e administradores")
    x = "Prof. Sec.",
    y = "Administrador"
  ) +
  tema +
  coord_cartesian(xlim = c(0,80000), ylim = c(0,80000))
```

Figura 4.3: Quantis entre o salário de professores secundaristas e administradore



O gráfico sugere que os salários de administradores são bastante superiores aos salários dos professores secundaristas.

Exercício 4.4

Para os dados do arquivo `salarios.xls`, considere a variável `Região`, com as classes `América do Norte`, `América Latina`, `Europa` e `Outros` e a variável `Salário de professor secundário`. Avalie a associação entre essas duas variáveis.

Solução. Iniciamos com a carga dos dados:

```
salarios <- statds_read("salarios.xls", "xls", skip = 3)
```

Agora vamos adicionar uma nova variável contendo a região. Começamos criando um vetor com os fatores que utilizaremos.

```
continente <- c("América do Norte", "América Latina", "Europa", "Outros")

regioes <- c("Europa", "Europa", "América Latina", "Europa", "América Latina", "América Latina", "América do Nor

salarios$`Região` <- parse_factor(regioes, levels = continente)
```

Como trata-se de uma variável contínua e outra categorica, vamos construir uma tabela resumindo os salários por continente e também exibir um diagrama de *box-plots*.

```
salarios %>%
  statds_summarise(`Prof. Sec.`, `Região`) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.5:** Medidas de resumo para a variável `Salário de Professor Secundarista` conforme `R
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

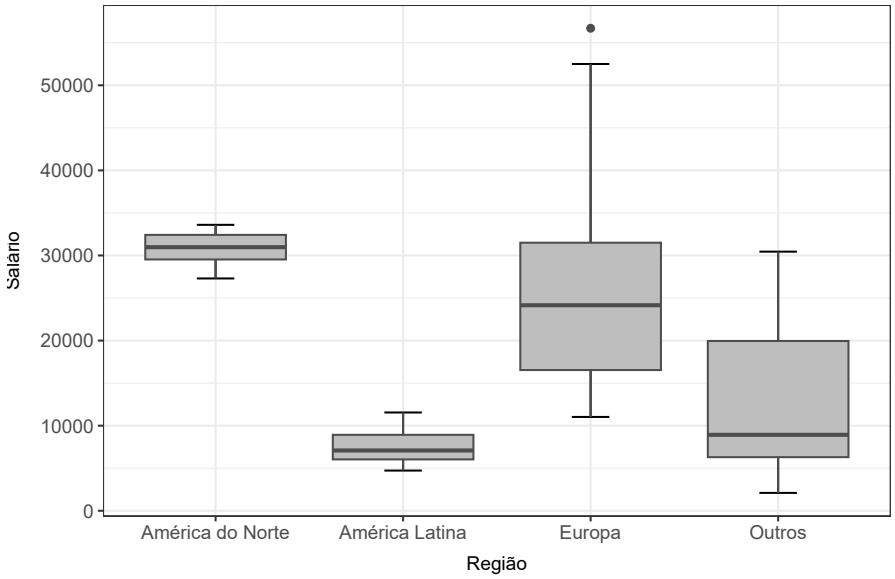
Tabela 4.5: Tabela 4.5: Medidas de resumo para a variável `Salário de Professor Secundarista` conforme `Região`

Região	n	Média	Variância	Desvio Padrão	Min.	Q1	Mediana	Q3	Máx.	IQR
América do Norte	6	30800,00	5475750	2340,03	27300	29531,25	30975,00	32418,75	33600	2887,5
América Latina	6	7612,50	6256688	2501,34	4725	6037,50	7087,50	8925,00	11550	2887,5
Europa	11	27467,73	232152600	15236,55	11025	16537,50	24150,00	31500,00	56700	14962,5
Outros	7	13425,00	12852000	3586,67	2100	6300,00	8925,00	19950,00	30450	13650,0

```
salarios %>%
  ggplot(aes(`Região`, `Prof. Sec.`)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
  labs(
    title = bquote(bold("Figura 4.4:")~"Salário de professor secundarista por região"),
    y = "Salário",
```

```
    x = "Região"
  ) +
  tema
```

Figura 4.4: Salário de professor secundarista por região



O cálculo de R^2 é dado por

```
r2 <- statds_rsquared(salarios, `Prof. Sec.`, `Região`)
```

Concluímos que o ganho de variância dado pela adição da variável `Região` é 0.5233435, ou seja, aproximadamente 52.33% da variação dos salários dos professores secundaristas é explicada pela região.

Exercício 4.5

Analise a variável `Preço de veículos` segundo as categorias `N` (nacional) e `I` (importado) para o conjunto de dados disponível no arquivo `veiculos.xls`.

Solução. Iniciaremos carregando os dados:

```
veiculos <- statds_read("veiculos.xls", "xls")
```

Novamente, utilizaremos uma tabela resumo, um gráfico de *boxplots* e o ganho de variância para avaliar visualmente a relação entre as variáveis.

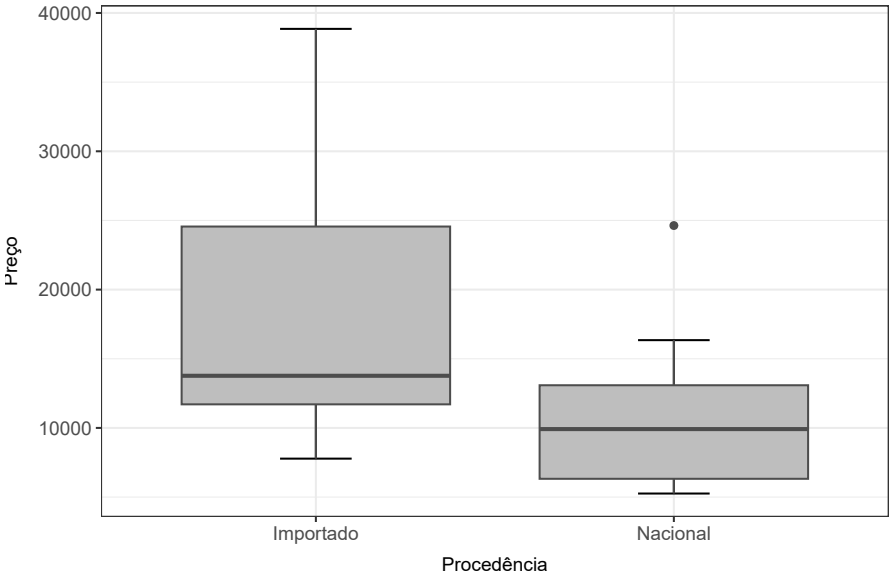
```
veiculos %>%
  statds_summarise(preco, proc) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.6:** Medidas de resumo para a variável `Preço do veículo` conforme a `Procedência`",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.6: Tabela 4.6: Medidas de resumo para a variável `Preço do veículo` conforme a `Procedência`

proc	n	Média	Variância	Desvio Padrão	Min.	Q1	Mediana	Q3	Máx.	IQR
Importado	12	19028,58	110250194	10500,01	7780	11705,25	13770	24560,00	38850	12854,75
Nacional	18	10574,44	25604052	5060,04	5257	6322,00	9916	13085,75	24632	6763,75

```
veiculos %>%
  ggplot(aes(proc, preco)) +
    stat_boxplot(
      geom = "errorbar",
      width = 0.2
    ) +
    geom_boxplot(
      fill = "grey",
      color = "grey30"
    ) +
  labs(
    title = bquote(bold("Figura 4.5:")~"Preço dos veículos conforme a procedência"),
    y = "Preço",
    x = "Procedência"
  ) +
  tema
```

Figura 4.5: Preço dos veículos conforme a procedência



Apesar de termos uma diferença nos preços, não podemos afirmar que há uma relação entre as variáveis. O coeficiente de determinação R^2 é 0.0891215, indicando que apenas aproximadamente 9% da variação de preço pode ser explicada pela procedência do veículo.

Exercício 4.6

Considere o conjunto de dados disponível no arquivo `coronarias.xls` [Singer and Ikeda, 1996].

a. Construa gráficos QQ para comparar a distribuição da variável COL de pacientes masculinos (= 1) com aquela de femininos (= 0). Repita a análise para a variável IMC e discuta os resultados.
b. Calcule o coeficiente de correlação de Pearson e o coeficiente de correlação de Spearman entre as variáveis ALTURA e PESO.

c) Construa uma tabela de contingência para avaliar a distribuição conjunta das variáveis TABAG4 e ARTER e calcule a intensidade de associação entre elas utilizando a estatística de Pearson, o coeficiente de contingencia de Pearson e o coeficiente de Tschuprov.

Solução. Iniciaremos carregando os dados.

```
coronarias <- statds_read("coronarias.xls", "xls", sheet = "dados")
coronarias <- coronarias %>%
                mutate(
                  IMC = parse_double(IMC),
                  PESO = parse_double(PESO),
                  ALTURA = parse_double(ALTURA),
                  ARTER = as.factor(ARTER),
                  TABAG4 = as.factor(TABAG4)
                 )

print(head(coronarias))
```

```
## # A tibble: 6 x 70
##   IDENT    HA   COL   HDL   LDL  VLDL  TRIG  DIAB  GLIC    AH TABAG TABAG4
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct>
## 1     1     1   202    NA    NA    NA   171     0    99     1     0 0
## 2     2     0   225    49   142    NA   167     1   317     0     1 3
## 3     3     1    NA    NA    NA    NA    NA    NA    NA     2     1 1
## 4     4     0   213    31   168    14   136     0    75     0    NA <NA>
## 5     5     0   258    40   200    18    92     0    78     2     0 0
## 6     6     0   184    NA    NA    NA   110     0    92     0     0 0
## # i 58 more variables: IDADE_M <dbl>, ALTURA <dbl>, PESO <dbl>, IMC <dbl>,
## #   ASS <dbl>, ANGEST <dbl>, ANGINS <dbl>, IMP <dbl>, ICC <dbl>, IMA <dbl>,
## #   ARRIT <dbl>, ARTER <fct>, CAT <dbl>, CD <dbl>, DA <dbl>, DI <dbl>,
## #   CX <dbl>, NUMAL <dbl>, PLAQ <dbl>, TE <dbl>, FIB <dbl>, IDADE1 <dbl>,
## #   SEXO <dbl>, M_PA <dbl>, M_C <dbl>, PSR <dbl>, PDR <dbl>, IDA55 <dbl>,
## #   SEID <dbl>, LO3 <dbl>, OBESO <dbl>, IDREAL <dbl>, SELO <dbl>, AH2 <dbl>,
## #   AH3 <dbl>, `C/H` <dbl>, `L/H` <dbl>, IDA60 <dbl>, SEID6 <dbl>, ...
```

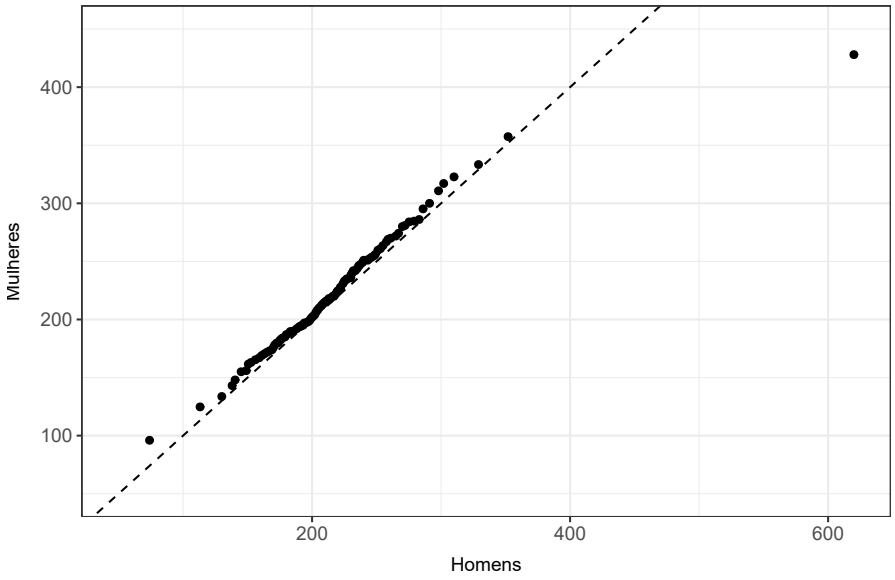
Colesterol de pacientes conforme o sexo

A figura a seguir mostra o gráfico QQ comparando os níveis de colesterol entre homens e mulheres.

```
quantiles <- seq(0, 1, 0.01)
homens <- coronarias %>% filter(SEXO == 1)
mulheres <- coronarias %>% filter(SEXO == 0)

q_col_homens <- quantile(homens$COL, quantiles, na.rm = TRUE)
q_col_mulheres <- quantile(mulheres$COL, quantiles, na.rm = TRUE)

ggplot(
```

```
  mapping = aes(
              x = q_col_homens,
              y = q_col_mulheres,
              slope = 1,
              intercept = 0
            )
) +
  geom_point() +
  geom_abline(linetype = 2) +
  labs(
    title = bquote(bold("Figura 4.6:")~"Quantis dos níveis de colesterol entre homens e mulheres"),
    x = "Homens",
    y = "Mulheres"
  ) +
  tema +
  coord_cartesian(xlim = c(50,620), ylim = c(50,450))
```

Figura 4.6: Quantis dos níveis de colesterol entre homens e mulheres



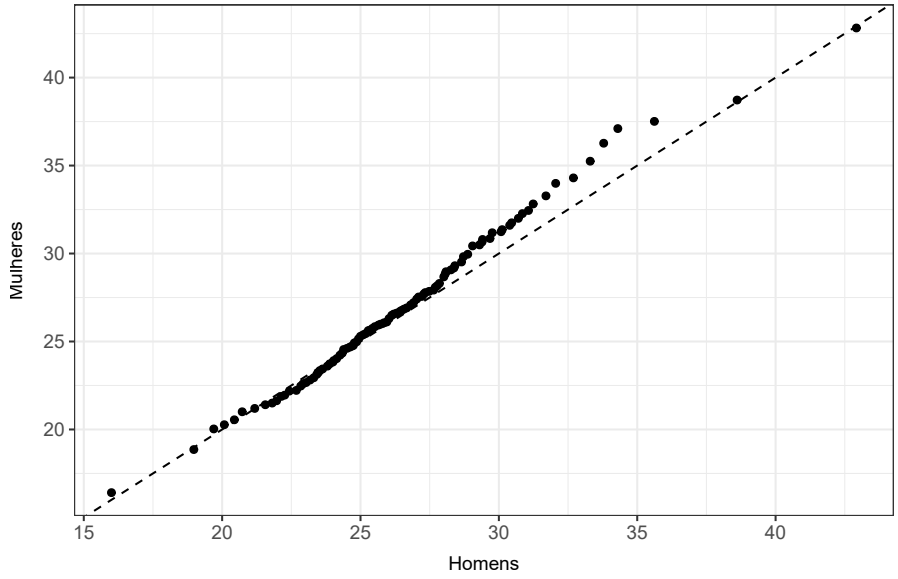
Como os pontos estão muito próximos da reta $y = x$, o gráfico sugere que há pouca diferença entre os níveis de colesterol dos dois grupos. O mesmo comportamento pode ser visto no gráfico de médias e diferenças a seguir.

IMc de pacientes conforme o sexo

A figura a seguir mostra o gráfico QQ comparando o IMC entre homens e mulheres.

```
q_imc_homens <- quantile(homens$IMC, quantiles, na.rm = TRUE)
q_imc_mulheres <- quantile(mulheres$IMC, quantiles, na.rm = TRUE)

ggplot(
  mapping = aes(
              x = q_imc_homens,
              y = q_imc_mulheres,
              slope = 1,
              intercept = 0
            )
) +
  geom_point() +
  geom_abline(linetype = 2) +
  labs(
    title = bquote(bold("Figura 4.7:")~"Quantis das medidas de IMC entre homens e mulheres"),
    x = "Homens",
    y = "Mulheres"
  ) +
  tema
```

Figura 4.7: Quantis das medidas de IMC entre homens e mulheres



O gráfico sugere que temos pouca diferença entre o IMC de homens de o de mulheres.

Correlação entre altura e peso

Para cálculo do coeficiente de correlação utilizaremos a função `cor()` do pacote base do R.

A expressão `cor(coronarias$ALTURA, coronarias$PESO, method = "pearson", use = "complete.obs")` calcula o coeficiente de correlação de Pearson e nos resulta o valor 0.5484947.

A expressão `cor(coronarias$ALTURA, coronarias$PESO, method = "spearman", use = "complete.obs")` calcula o coeficiente de correlação de Spearman e nos resulta o valor 0.559371.

Relação entre as variáveis TABAG4 e ARTER

```
coronarias %>%
  filter(!is.na(TABAG4)) %>%
  filter(!is.na(ARTER)) %>%
  group_by(TABAG4, ARTER) %>%
  count() %>%
  bind_rows(
    group_by(., TABAG4) %>%
      summarise(n = sum(n)) %>%
      mutate(ARTER = "Total")
  ) %>%
  bind_rows(
    group_by(., ARTER) %>%
      summarise(n = sum(n)) %>%
      mutate(TABAG4 = "Total")
  ) %>%
  pivot_wider(
    names_from = ARTER,
    values_from = n,
    values_fill = 0
  ) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.7:** Frequência de artereopatia entre quatro categorias de tabagismo",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.7: Tabela 4.7: Frequência de artereopatia entre quatro categorias de tabagismo

TABAG4	0	1	2	3	Total
0	457	43	27	2	529
1	210	11	15	1	237
2	259	19	5	2	285
3	257	15	15	3	290
4	8	0	1	0	9
5	2	0	0	0	2
Total	1193	88	63	8	1352

O trecho de código a seguir calcula as nossas estatísticas de interesse.

```
chisq <- chisq.test(coronarias$TABAG4, coronarias$ARTER)
```

```
## Warning in chisq.test(coronarias$TABAG4, coronarias$ARTER): Aproximação do
## qui-quadrado pode estar incorreta
```

```
chisq <- chisq$statistic[[1]]

n <- nrow(coronarias)
r <- 6
c <- 4

C <- sqrt(chisq / (chisq + n))
t <- sqrt((chisq / n) / ((r-1) * (c-1)))
```

A estatística de Pearson é 14.8672858, o coeficiente de contingência de Pearson é 0.0990669 e o coeficiente de Tschuprov é 0.0257054. Todas essas estatísticas nos sugerem que a associação entre as variáveis é muito fraca ou inexistente.

Exercício 4.7

Considere os dados do arquivo `endometriose.xls` [Abrão et al., 1997]. Construa um gráfico QQ para comparar as distribuições da variável `Idade` de pacientes dos grupos `Controle` e `Doente`.

Solução. No Exercício 3.12, já havíamos tratado os dados deste conjunto. Vamos carregar o arquivo final que geramos, que já contem os dados ajustados.

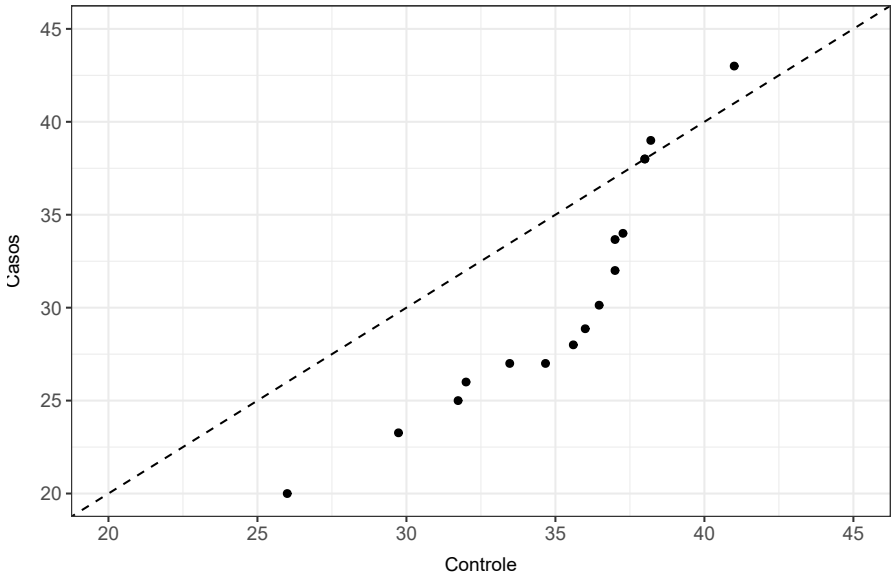
```
endometriose <- statds_read("endometriose.csv")
```

```
## Rows: 50 Columns: 13
## -- Column specification --------------------------------------------------------
---
## Delimiter: ","
## chr  (3): Grupo, Dismenorréia, Dispareunia
## dbl (10): Paciente, Idade, Gestação, Partos, Abortos, AFSr, CA125/A, CA125/B...
##
## 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.
```

Com os dados carregados, vamos contruir o gráfico QQ:

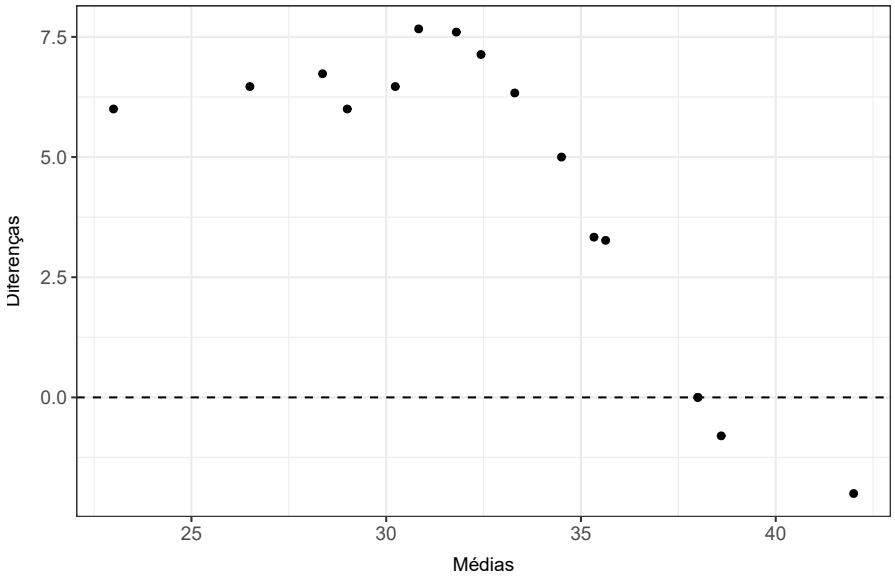
```
controle <- (endometriose %>% filter(Grupo == "Controle"))$Idade
doente <- (endometriose %>% filter(Grupo == "Doente"))$Idade

quantiles <- seq(0, 1, 1/15)
q_controle <- quantile(controle, quantiles)
q_doente <- quantile(doente, quantiles)

ggplot() +
  geom_point(aes(q_controle, q_doente)) +
  geom_abline(aes(slope = 1,intercept = 0), linetype = 2) +
  labs(
    title = bquote(bold("Figura 4.8:")~"Quantis das idades dos pacientes dos grupos Caso e Controle"),
    x = "Controle",
    y = "Casos"
  ) +
  tema +
  coord_cartesian(
    xlim = c(20,45),
    ylim = c(20,45)
  )
```

Figura 4.8: Quantis das idades dos pacientes dos grupos Caso e Controle




```
q_media <- (q_controle + q_doente) / 2
q_diferenca <- q_controle - q_doente

ggplot() +
  geom_point(aes(q_media, q_diferenca)) +
  geom_hline(yintercept = 0, linetype = 2) +
  labs(
    title = bquote(bold("Figura 4.9:")~"Médias"~italic("versus")~"diferenças entre os quantis das idades das pac
    x = "Médias",
    y = "Diferenças"
  ) +
  tema
```

Figura 4.9: Médias *versus* diferenças entre os quantis das idades das pacientes



O gráfico da Figura 4.8 nos mostra que existe uma diferença entre as idades das pacientes, sendo as do grupo controle mais jovens do que as do grupo de casos. A Figura 4.9, por sua vez, nos dá uma visão mais clara da magnitude dessa diferença (diferença média de 4.325 anos e desvio padrão de 3.2725005 anos).

Exercício 4.8

Considere os dados do arquivo `neonatos.xls` contendo pesos de recém nascidos medidos por via ultrassonográfica (antes do parto) e ao nascer. Construa gráficos QQ e gráficos Bland-Altman para avaliar a concordância entre as duas distribuições. Comente os resultados.

Solução.

Carregamento dos dados

```
neonatos <- statds_read("neonatos.xls", "xls")
```

Gráfico QQ

```
quantiles <- seq(0, 1, 1/68)
q_peso1 <- quantile(neonatos$peso1, quantiles)
q_peso2 <- quantile(neonatos$peso2, quantiles)

ggplot() +
  geom_point(aes(q_peso1, q_peso2)) +
  geom_abline(aes(slope = 1, intercept = 0), linetype = 2) +
  labs(
    title = bquote(bold("Figura 4.10:")~"Quantis dos pesos de recém nascidos"),
    x = "Peso obtido ultrassonograficamente (g)",
    y = "Peso ao nascer (g)"
  ) +
  tema
```

Figura 4.10: Quantis dos pesos de recém nascidos

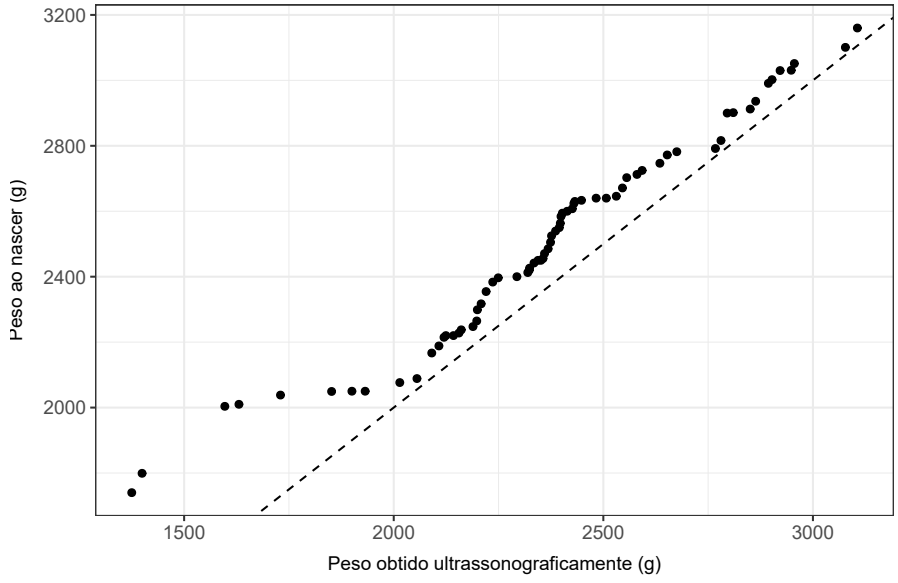
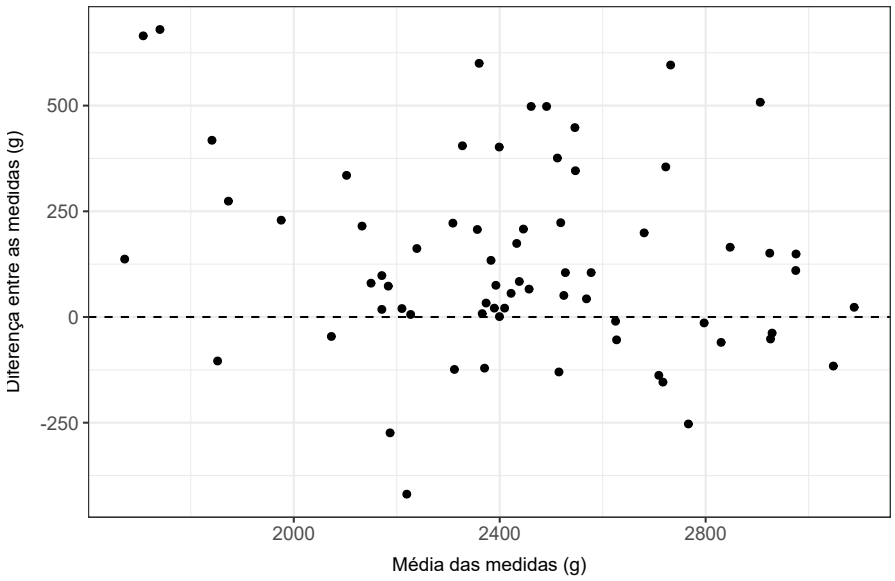


Gráfico Bland-Altman

```
q_media <- (neonatos$peso2 + neonatos$peso1) / 2
q_diferenca <- (neonatos$peso2 - neonatos$peso1)

ggplot() +
  geom_point(aes(q_media, q_diferenca)) +
  geom_hline(yintercept = 0, linetype = 2) +
  labs(
    title = bquote(bold("Figura 4.11:")~"Diferença entre as medidas do peso dos recém nascidos"),
    x = "Média das medidas (g)",
    y = "Diferença entre as medidas (g)"
  ) +
  tema
```

Figura 4.11: Diferença entre as medidas do peso dos recém nascidos



Comentários

A Figura 4.10 nos mostra que a medida do peso obtida por via ultrassonográfica é levemente menor do que aquela obtida ao nascer. A Figura 4.11 nos dá uma melhor visão dessa diferença (média: 131.8970588g, desvio padrão: 227.9162677g, mediana: 91g). A média das diferenças corresponde a um percentual de 5.27 em relação ao peso real aferido ao nascer.

Exercício 4.9

Considere o conjunto de dados disponível no arquivo `esforco.xls` [**?**].

a. Compare as distribuições de frequências da variável `V02` em repouso e no pico do exercício para pacientes classificados em cada um dos níveis da variável `Etiologia` por meio de gráficos QQ e de medidas resumo. Comente os resultados.
b. Repita o item (a) utilizando gráficos de Bland-Altman.
c. Utilize *boxplots* e gráficos de perfis médios para comparar as distribuições da variável `FC` correspondentes a pacientes nos diferentes níveis da variável `NYHA`. Comente os resultados.

Solução.

Carregamento dos dados

```
esforco <- statds_read("esforco_completo.csv")
```

```
## Rows: 127 Columns: 61
## -- Column specification --------------------------------------------------------
---
## Delimiter: ","
## chr   (7): iniciais, etiologia, sexo, weber, teste_maximo_rep, cirurgia, grupo
## dbl  (51): id_grupo, idade, altura, peso, superficie_corporal, imc, nyha, fc...
## date  (3): data_espirometrico, data_obito, data_ultima_obs
##
## 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.
```

4.6.0.1 Gráfico QQ para o consumo de oxigênio

As tabelas abaixo apresentam medidas de resumo para o consumo de oxigênio por grupo.

```
esforco %>%
  statds_summarise(vo2_rep, etiologia) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.8:** Medidas de resumo para o consumo de oxigênio em repouso",
```

```
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.8: Tabela 4.8: Medidas de resumo para o consumo de oxigênio em repouso

etiologia	n	Média	Variância	Desvio Padrão	Min.	Q1	Mediana	Q3	Máx.	IQR
Chagasíaca	26	3,75	0,91	0,96	2,6	3,02	3,45	4,32	5,9	1,30
Controle	40	3,50	0,36	0,60	2,4	3,00	3,45	3,85	5,1	0,85
Idiopático	31	3,66	1,00	1,00	2,4	3,10	3,50	3,90	6,4	0,80
Isquêmico	30	3,35	0,54	0,73	1,7	3,00	3,25	3,88	4,9	0,88

```
esforco %>%
  statds_summarise(vo2_w, etiologia) %>%
  kable(
    format = "pipe",
    caption = "**Tabela 4.9:** Medidas de resumo para o consumo de oxigênio no pico do exercício",
    label = NA,
    digits = 2,
    align = "c",
    format.args = list(decimal.mark = ",")
  )
```

Tabela 4.9: Tabela 4.9: Medidas de resumo para o consumo de oxigênio no pico do exercício

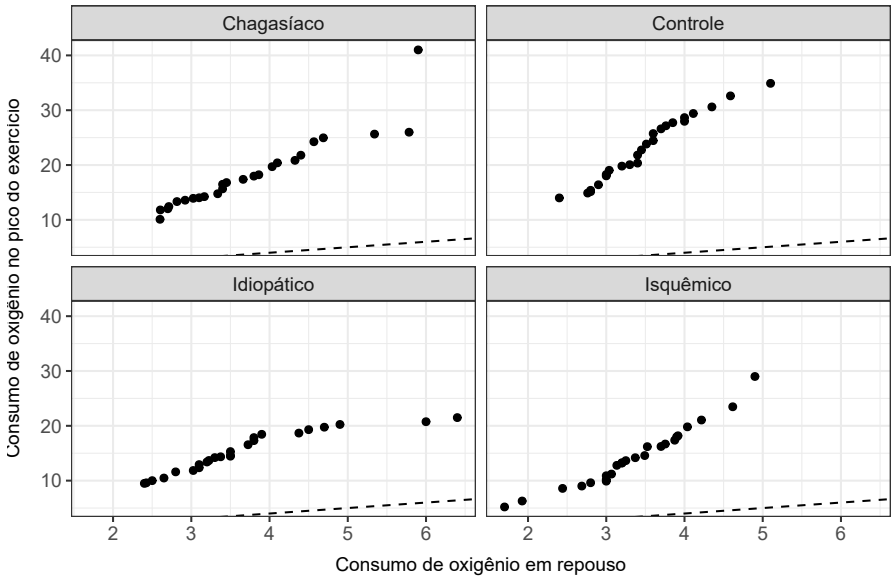
etiologia	n	Média	Variância	Desvio Padrão	Min.	Q1	Mediana	Q3	Máx.	IQR
Chagasíaca	26	18,27	42,66	6,53	10,1	13,93	16,80	20,85	41,0	6,93
Controle	40	23,00	33,89	5,82	14,0	18,28	22,75	27,72	34,9	9,45
Idiopático	31	15,13	13,41	3,66	9,5	12,35	14,46	18,45	21,5	6,10
Isquêmico	30	14,30	28,50	5,34	5,2	10,20	13,65	17,38	29,0	7,18

Para construção dos gráficos, vamos precisar fazer algumas adaptações:

```
quantiles <- seq(0, 1, 1/24)

qq_rep <- function(x) quantile(x$vo2_rep, quantiles)
qq_w <- function(x) quantile(x$vo2_w, quantiles)

qq_esforco <- esforco %>%
                select(etiologia, vo2_rep, vo2_w) %>%
                mutate(
                  qq_media = (vo2_w + vo2_rep) / 2,
                  qq_diff = vo2_w - vo2_rep
                ) %>%
                group_by(etiologia) %>%
                nest() %>%
                mutate(
                  qq_rep = map(data, qq_rep),
                  qq_w = map(data, qq_w),
                )

qq_esforco %>%
  unnest(qq_rep, qq_w) %>%
  ggplot(aes(qq_rep, qq_w, slope = 1, intercept = 0)) +
    geom_point() +
    geom_abline(linetype = 2) +
    facet_wrap(~etiologia) +
    labs(
      title = bquote(bold("Figura 4.12:")~"Quantis do consumo de oxigênio em repouso"~italic("vs.")~"pico do exe
      x = "Consumo de oxigênio em repouso",
      y = "Consumo de oxigênio no pico do exercício"
    ) +
    tema
```

```
## Warning: `unnest()` has a new interface. See `?unnest` for details.
## i Try `df %>% unnest(c(qq_rep, qq_w))`, with `mutate()` if needed.
```

Figura 4.12: Quantis do consumo de oxigênio em repouso *vs.* pico do exercício

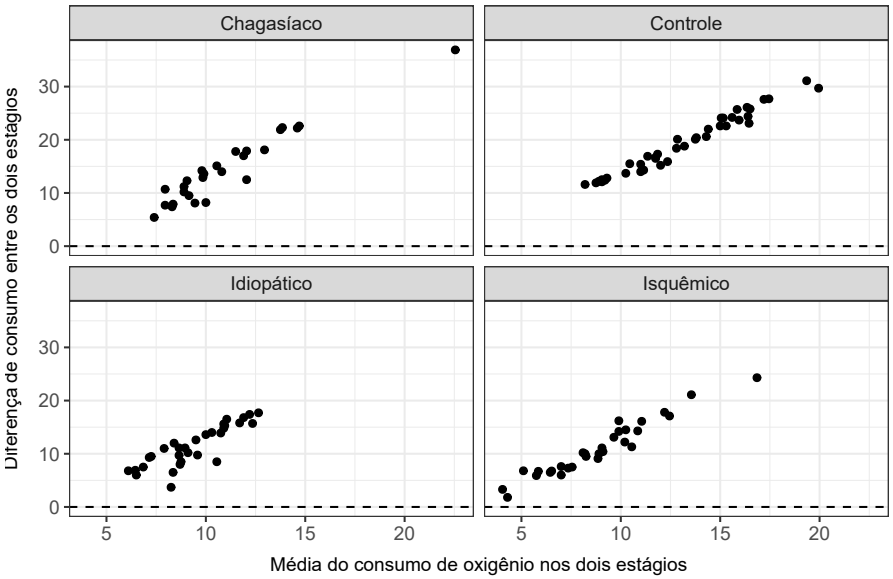


O gráfico nos mostra uma grande diferença entre os dados, sugerindo que o consumo de oxigênio no pico do exercício é substancialmente maior do que quando em repouso.

Gráfico de Bland-Altman para o consumo de oxigênio

```
qq_esforco %>%
  unnest(data) %>%
  ggplot(aes(qq_media, qq_diff)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = 2) +
    facet_wrap(~etiologia) +
    labs(
      title = bquote(bold("Figura 4.13:")~"Média do consumo de oxigênio"~italic("vs.")~"diferença entre os estág
      x = "Média do consumo de oxigênio nos dois estágios",
      y = "Diferença de consumo entre os dois estágios"
    ) +
    tema
```

Figura 4.13: Média do consumo de oxigênio *vs.* diferença entre os estágios



Avaliar melhor.

Exercício 4.10

Os dados da Tabela 4.29 são provenientes de um estudo em que um dos objetivos era avaliar o efeito da dose de radiação gama (em centigrays) na formação de múltiplos micronúcleos em células de individuos normais. Analise os dados descritivamente, calculando o risco relativo de ocorrência de micronúcleos para cada dose tomando como base a dose nula. Repita a análise calculando as razões de chances correspondentes. Quais as conclusões de suas análises?

Tabela 4.10: Tabela 4.29: Número de células

Dose de radiação gama (cGy)	Frequência de células com múltiplos micronúcleos	Total de células examinadas
0	1	2373
20	6	2662
50	25	1991
100	47	2047
200	82	2611
300	207	2442
400	254	2398
500	285	1746

Solução. x

Exercício 4.11

Numa cidade A em que não foi veiculada propaganda, a porcentagem de clientes que desistem do plano de TV a cabo depois de um ano é 14%. Numa cidade B, em que houve uma campanha publicitária, essa porcentagem é de 6%. Considerando uma aproximação de 2 casas decimais, indique qual é a razão de chances (re) de desistência entre as cidades A e B, justificando sua resposta:

a) $re = 2.33$
b) $re = 2.55$
c) $re = 8.00$
d) $re = 1.75$
e) Nenhuma das respostas anteriores está correta.

Solução. x

Exercício 4.12

De uma tabela construída para avaliar a associação entre tratamento (com níveis ativo e placebo) e cura (sim ou não) de uma certa moléstia obteve-se uma razão de chances igual a 2,0. Mostre que não se pode concluir dai que a probabilidade de cura para pacientes submetidos ao tratamento ativo é 2 vezes a probabilidade de cura para pacientes submetidos ao placebo.

Solução. x

Exercício 4.13

Considere os dados do arquivo `esquistossomose.xls`. Calcule a sensibilidade, especificidade, taxas de falsos positivos e falsos negativos, valores preditivos positivos e negativos e acurácia correspondentes aos cinco testes empregados para diagnóstico de esquistossomose.

Solução. x

Exercício 4.14

Considere os dados do arquivo `entrevista.xls`. Calcule estatísticas κ sem e com ponderação para quantificar a concordância entre as duas observadoras (G e P) para as variáveis `Impacto` e `Independência` e comente os resultados.

Solução. x

Exercício 4.15

Considere os dados do arquivo `figadodiag.xls`. Calcule a sensibilidade, especificidade, taxas de falsos positivos e falsos negativos, valores preditivos positivos e negativos e acurácia das técnicas radiológicas para detecção de alterações anatômicas na veia porta e na via biliar tendo os resultados intraoperatórios como padrão ouro.

Solução. x

Exercício 4.16

Um criminologista desejava estudar a relação entre: X (densidade populacional = número de pessoas por unidade de área) e Y (índice de assaltos = número de assaltos por 100000 pessoas) em grandes cidades. Para isto sorteou 10 cidades observando em cada uma delas os valores de X e Y. Os resultados obtidos estão dispostos na Tabela 4.30

Tabela 4.11: Tabela 4.30: Densidade populacional e índice de assaltos em grandes cidades

Cidade	1	2	3	4	5	6	7	8	9	10
X	59	49	75	65	89	70	54	78	56	60
Y	190	180	198	186	200	204	192	215	197	208

a. Classifique as variáveis envolvidas.
b. Calcule a média, mediana, desvio padrão e a distância interquartis para cada varikvel
c. Construa o diagrama de dispersão entre Ye Xe faça comentários sobre a relação entre as duas variáveis.

Solução. x

Exercício 4.17

Considere a seguinte tabela.

X | 1 | 3 | 4 | 6 | 8 | 9 | 11 | 14 |
Y | 1 | 2 | 4 | 4 | 5 | 7 | 8 | 9 |

Indique qual a afirmação abaixo sobre a relação entre as variáveis X e Y é correta, justificando sua resposta.

a) Não há associação entre X e Y.
b) Há relação linear positiva.

c) Hå relação linear negativa.
d) Há relação quadrática.

Solução. x

Exercício 4.18

Em um teste de esforço cardiopulmonar aplicado a 55 mulheres e 104 homens, foram medida entre outras, as seguintes variáveis:

- Grupo: Normais, Cardiopatas ou DPOC (portadores de doença pulmonar obstrutiva crônica).
- VO2MAX: consumo máximo de O_2 (ml/min).
- VCO2MAX: consumo máximo de CO_2 (ml/min).

Algumas medidas descritivas e gráficos são apresentados abaixo nas Tabelas 4.31 e 4.32 e Figura 4.21.

Tabela 4.12: Tabela 4.31: VO2MAX

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	1845	1707	795
Cardiopatas	57	1065	984	434
DPOC	46	889	820	381

Tabela 4.13: Tabela 4.32: VCO2MAX

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	2020	1847	918
Cardiopatas	57	1206	1081	479
DPOC	46	934	860	430

Coeficiente de correlação entre VO2MAX e VCO2MAX = 0,92.

a. Que grupo tem a maior variabilidade?
b. Compare as médias e as medianas dos 3 grupos.
c. Compare as distâncias interquartis dos 3 grupos para cada variável. Você acha rasi usar a distribuição normal para esse conjunto de dados?
d. O que representam os asteriscos nos boxplots?
e. Que tipo de função você ajustaria para modelar a relação entre o consumo máximo de CO_2 e o consumo máximo de O_2? Por quê?

f. Há informações que necessitam verificação quanto à possíveis erros? Quais?

Solução. x

Exercício 4.19

Para avaliar a associação entre a persistência do canal arterial (PCA) em recém-nascidos pré-termo (RNPT) e óbito ou hemorragia intracraniana (HI), um pesquisador obteve os dados dispostos na seguinte tabela de frequências:

PCA	Sim	Óbito Não	Total	Sim	HI Não	Total
Presente	8	13	21	7	14	21
Ausente	1	39	40	7	33	40
Total	9	52	61	14	44	61

Um resumo das análises para óbitos e hemorragia intracraniana está disposto na tabela seguinte:

a. Interprete as estimativas das razões de chances, indicando claramente a que pacientes elas se referem.
b. Analogamente, interprete os intervalos de confiança correspondentes, indicando claramente a que pacientes eles se referem.
c. Com base nos resultados anteriores, o que você pode concluir sobre a associação entre persistência do canal arterial e óbito para RNPT em geral? E sobre a associação entre a persistência do canal arterial e a ocorrência de hemorragia interna? Justifique suas respostas.
d. Qual a hipótese nula testada em cada caso?
e. Qual a interpretação dos valores-p em cada caso?

Detalhes podem ser obtidos em Afiune (2000).

Solução. x

Exercício 4.20

Em um estudo realizado para avaliar o efeito do tabagismo nos padrões de sono foram consideradas amostras de tamanhos 12 e 15 de duas populações: Fumantes e Não Fumantes, respectivamente. A variável observada foi o tempo, em minutos, que se leva para dormir. Os correspondentes *boxplots* e gráficos de probabilidade normal são apresentados nas Figuras 4.22 e 4.23.

Esses gráficos sugerem que:

a. a variabilidade do tempo é a mesma nas duas populações estudadas;
b. as suposições para a aplicação do teste t-Student para comparar as médias dos tempos nas duas populações estão válidas;
c. os fumantes tendem a apresentar um tempo maior para dormir do que os não fumantes;
d. as informações fornecidas permitem concluir que o estudo foi bem planejado;
e. nenhuma das respostas anteriores está correta.

Solução. x

Exercício 4.21

Em um estudo comparativo de duas drogas para hipertensão os resultados indicados nas Tabelas 4.33, 4.34 e 4.35 e Figura 4.24 foram usados para descrever a eficácia e a tolerabilidade das drogas ao longo de 5 meses de tratamento.

a. Com a finalidade de melhorar a apresentação dos resultados, faça as alterações que você julgar necessárias em cada uma das tabelas e figura.
b. Calcule a média, o desvio padrão e a mediana da variação de pressão arterial para cada uma das duas drogas por meio do histograma.
c. Compare os resultados obtidos no item b com aqueles obtidos diretamente dos dados da amostra (Tabela 4.35).

Solução. x

Exercício 4.22

Considere duas amostras de uma variável X com n unidades amostrais cada. Utilize a definição (4.9) para mostrar que Var(X) = Var(X) quando as médias das duas amostras são iguais.

Solução. x

Exercício 4.23

Utilize o método Delta para calcular uma estimativa da variancia da razão de chances (ver Nota de Capítulo 7).

Solução. x

Exercício 4.24

Utilizando a definição da Nota de Capítulo 4, prove que se $\alpha = 0$, então $r(\alpha) = r$.

Solução. x

Exercício 4.25

Mostre que para a hipótese de inexistência de associação numa tabela $r \times s$, a estatística (4.1) pode ser escrita como

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$$

em que n_{ij} é a frequência absoluta observada na linha i e coluna j e n_{i+} e n_{+j} são, respectivamente, os totais das linhas e colunas.

Solução. x

Exercício 4.26

Prove que a expressão da estatística de Pearson do Exercício 4.10 pode ser escrita como

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

,

em que :f_{ij} e f_{ij}^* representam, respectivamente, as frequências relativas observada e esperada (sob a hipótese de inexistência de associação) correspondentes à casela (i, j).

Solução. x

Exercício 4.27

Prove que (4.4) e (4.5) são equivalentes.

Solução. x

Exercício 4.28

Prove as relações (4.12)-(4.14).

Solução. x

5

Análise de dados de várias variáveis

6

Análise de Regressão

7

Análise de Sobrevivência

Parte II

Aprendizado Supervisionado

8

Regularização e Modelos Aditivos Generalizados

9

Classificação por meio de técnicas clássicas

10

Algoritmos de Suporte Vetorial

11

Árvores e Florestas

12

Redes neurais

Parte III

Aprendizado não Supervisionado

13

Análise de Agrupamentos

14

Redução de dimensionalidade

A

Otimização numérica

B

Noções de simulação

C

Algoritmos para dados aumentados

Bibliografia

Mauricio S. Abrão, Sergio Podgaec, Braz Martorelli Filho, Laudelino O. Ramos, José Aristodemo Pinotti, and Ricardo M. de Oliveira. The use of biochemical markers in the diagnosis of pelvic endometriosis. *Human Reproduction*, 12(11): 2523–2527, 11 1997. ISSN 0268-1161. doi: 10.1093/humrep/12.11.2523. URL https://doi.org/10.1093/humrep/12.11.2523.

Ana Maria Fonseca Wanderley Braga. *Comportamento e valor prognóstico das variáveis obtidas no teste de esforço cardiopulmonar em portadores de insuficiência cardíaca*. Tese de doutorado, Faculdade de Medicina da Universidade de São Paulo, São Paulo, 1998.

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.

Julio da Motta Singer and Karina Ikeda. Relatório de análise estatística sobre o projeto "fatores de risco na doença aterosclerótica coronariana", 1996. URL https://repositorio.usp.br/directbitstream/1c00ad75-fdc7-4281-9f3c-cd777c12515b/963591%20-%20Relat%C3%B3rio%20de%20an%C3%A1lise%20estat%C3%ADstica%20sobre%20o%20projeto%20fatores%20de%20risco%20na%20doen%C3%A7a%20ateroscler%C3%B3tica%20coronariana.pdf. RAECEA-9608.