



Análise Exploratória de Dados

Especialização em Estatística Aplicada

Prof. Diógenes Ferreira Filho

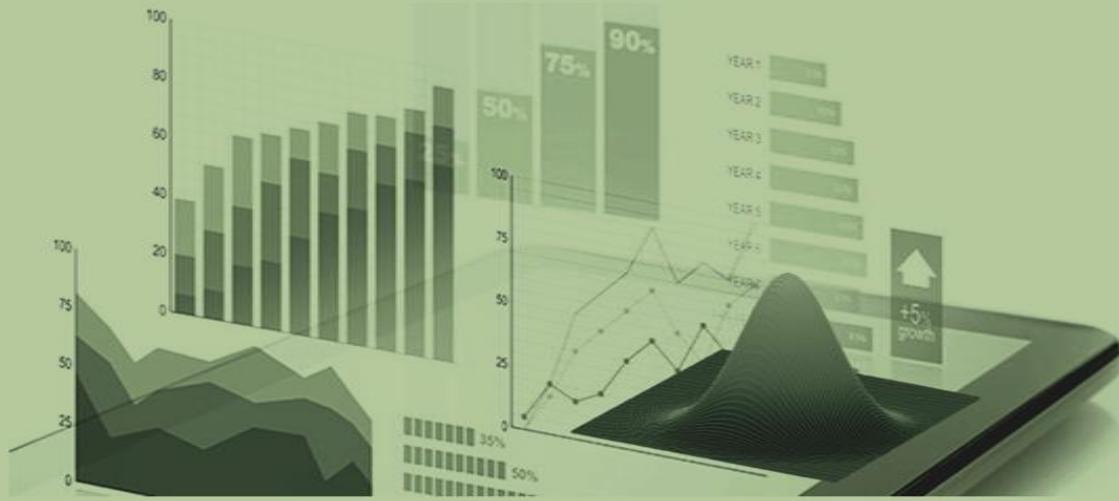
Prof.ª Adriana Oliveira Andrade

Prof. Dr. Diógenes Ferreira Filho
Departamento de Ciências Econômicas e Exatas / Instituto Três Rios / UFRRJ
e-mail: dffilho@gmail.com / dffilho@ufrj.br

Prof.^a Dra. Adriana Oliveira Andrade
Departamento de Matemática / Instituto de Ciências Exatas / UFRRJ
e-mail: andrade.ufrj@gmail.com

Esta apostila foi preparada para cobrir o programa da disciplina **Análise Exploratória de Dados** do curso de *Especialização em Estatística Aplicada* do Departamento de Matemática da Universidade Federal Rural do Rio de Janeiro. São abordados temas que vão desde a instalação e introdução à linguagem R até técnicas de resumo e apresentação de dados estatísticos tratando dos aspectos teórico e prático/computacional. A apostila é composta por notas de aula elaboradas a partir de livros e apostilas constantes na Bibliografia e não substitui a leitura dos mesmos. O material não está livre de erros e/ou imperfeições e toda e qualquer contribuição será bem-vinda.

Seropédica - RJ, agosto de 2020.



Sumário

I	Seção I	
1	Download e instalação do Software R	9
1.1	Sobre o R	9
1.2	Download e instalação do R	9
2	Download e instalação do RStudio	13
2.1	Sobre o RStudio	13
2.2	Download e instalação do RStudio	13
3	Introdução ao Software R	17
3.1	Console	17
3.2	Script	17
3.3	Pacotes	18
3.4	Operações matemáticas básicas	19
3.5	Armazenamento de dados	19
3.6	Operadores lógicos	20
3.7	Arquivos do R	21
3.8	Estrutura de dados	21
3.8.1	Vetores	21
3.8.2	Matrizes	24
3.8.3	Data-frame	25
3.8.4	Listas	26
3.9	Importação de conjunto de dados para o R	27

4	Tipos de variáveis	31
4.1	Variáveis quantitativas	31
4.2	Variáveis qualitativas	32
5	Tabelas	33
5.1	Normas para construção de tabelas	33
5.2	Tabelas de distribuição de frequências	35
5.2.1	Distribuição de frequências para dados qualitativos	35
5.2.2	Distribuição de frequências para dados quantitativos	37
6	Gráficos	43
6.1	Gráficos de colunas e barras	43
6.1.1	Gráfico de colunas (ou gráfico de barras verticais)	43
6.1.2	Gráfico de barras (ou gráfico de barras horizontais)	45
6.2	Gráfico de setores	46
6.3	Diagrama de linhas	48
6.4	Histograma	50
6.5	Polígono de frequências	52
6.6	Ramo e folhas	55
6.7	Gráfico de Série Temporal	57
7	Medidas de posição	61
7.1	Introdução	61
7.2	Somatório	61
7.2.1	Variáveis e índices	61
7.2.2	Notação de somatório	62
7.2.3	Propriedades de somatório	63
7.3	Média aritmética	64
7.3.1	Média aritmética (para dados brutos)	64
7.3.2	Média aritmética (para dados agrupados)	65
7.4	Mediana	68
7.4.1	Mediana para dados brutos	68
7.4.2	Mediana para dados agrupados	69
7.5	Moda	72
7.5.1	Moda para dados brutos	72
7.5.2	Moda para dados agrupados	73
7.5.3	Moda para dados qualitativos	76
7.6	Utilização das medidas de tendência central	76
7.7	Medidas separatrizes (quartis)	77
7.7.1	Quartis para dados brutos	77
7.7.2	Quartis para dados agrupados	80
7.7.3	Percentis	82

7.8	Boxplot	83
7.8.1	Boxplot comparativo	85
8	Medidas de dispersão	89
8.1	Amplitude	89
8.2	Desvio médio absoluto	90
8.3	Variância	92
8.3.1	Variância para dados brutos	92
8.3.2	Variância para dados agrupados	94
8.4	Desvio padrão	96
8.5	Coeficiente de variação	96
8.6	Propriedades das Medidas de Dispersão	99
9	Assimetria e curtose	101
9.1	Assimetria	101
9.2	Curtose	105
10	Análise Bivariada	109
10.1	Variáveis qualitativas	109
10.1.1	Tabelas de contingência e gráficos da distribuição conjunta	109
10.1.2	Associação entre Variáveis Qualitativas	111
10.1.3	Qui-quadrado (χ^2)	114
10.1.4	Coeficiente de contingência	116
10.2	Variáveis quantitativas	117
10.2.1	Gráfico de dispersão	117
10.2.2	Coeficiente de correlação	119
	Bibliografia	123

Seção I

1	Download e instalação do Software R	9
1.1	Sobre o R	
1.2	Download e instalação do R	
2	Download e instalação do RStudio	13
2.1	Sobre o RStudio	
2.2	Download e instalação do RStudio	
3	Introdução ao Software R	17
3.1	Console	
3.2	Script	
3.3	Pacotes	
3.4	Operações matemáticas básicas	
3.5	Armazenamento de dados	
3.6	Operadores lógicos	
3.7	Arquivos do R	
3.8	Estrutura de dados	
3.9	Importação de conjunto de dados para o R	



1. Download e instalação do Software R

1.1 Sobre o R

R é uma linguagem de programação e também um ambiente computacional e estatístico. O software R fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento, ...), técnicas gráficas e é altamente expansível, o que amplia sua aplicabilidade. O R está disponível gratuitamente como Software Livre e roda em uma ampla variedade de plataformas como Linux, Windows e MacOS. Maiores informações em <https://www.r-project.org/about.html>

1.2 Download e instalação do R

O R pode ser obtido gratuitamente no endereço <https://www.r-project.org>. Estão disponíveis versões para os sistemas operacionais Windows, MacOS e Linux. Após acessar o site, deve-se escolher o CRAN (Comprehensive R Archive Network) para indicar o local de disponibilização do software. No Brasil, atualmente, temos cinco opções: Universidade Estadual de Santa Cruz, Universidade Federal do Paraná, Fundação Oswaldo Cruz, Universidade de São Paulo e USP de Piracicaba.

Neste material serão apresentados os passos para download e instalação do R apenas para o sistema operacional Windows. Na Figura 1.1 podemos ver os passos para fazer o download do software R. Na página inicial do site do R (Figura 1.1 A) clique em "CRAN" e você será redirecionado para outra página (Figura 1.1 B). Nesta página clique em um dos repositórios do Brasil (Fiocruz, por exemplo) e você será redirecionado para outra página (Figura 1.1 C). Nesta página clique em "Download R for Windows" e você será redirecionado para outra página (Figura 1.1 D). Nesta página clique em "base" e você será redirecionado para outra página (Figura 1.1 E) onde se encontra a versão mais recente do R. Nesta página clique em "Download R 4.0.0 for Windows" (até o momento em que este material foi elaborado a versão mais recente do R era a 4.0.0) e será feito o download do arquivo de instalação do R (Figura 1.1 F).

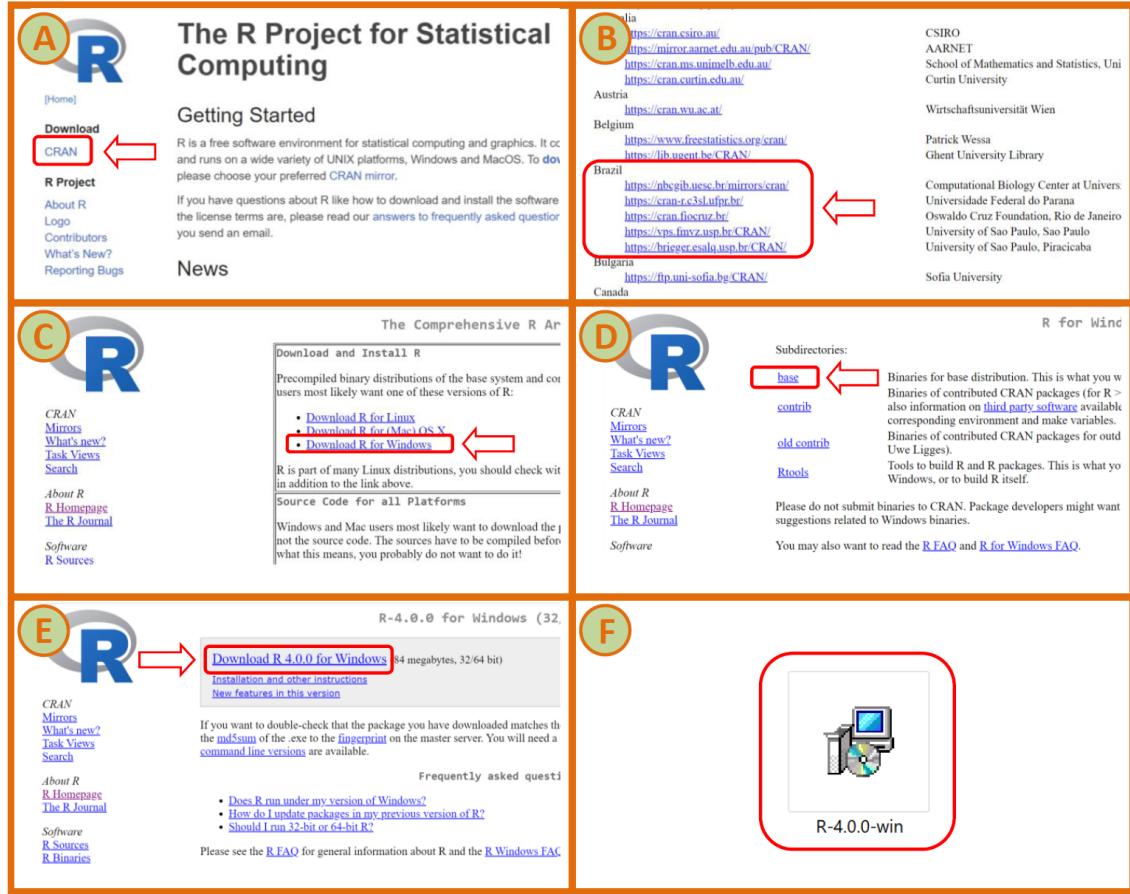


Figura 1.1: Passos para fazer o download do R

Após fazer o download do arquivo de instalação do R (Figura 1.1 F) dê um clique duplo sobre ele para executá-lo. Selecione o idioma "Português Brasileiro" (Figura 1.2 A) e clique em "OK". Nas próximas janelas que irão abrir (Figuras 1.2 B à G) basta clicar em "Próximo" para fazer a instalação padrão do R. Por fim clique em "Concluir" (Figura 1.2 H) para finalizar a instalação.

Após finalizar a instalação será criado um atalho para o software R na área de trabalho. Dê um clique duplo sobre o atalho para executar o R.

Na Figura 1.3 pode-se observar a aparência do software R. Ao abrir o programa é mostrado um menu superior e uma janela principal chamada *R Console*. É nesta janela que os comandos são executados.

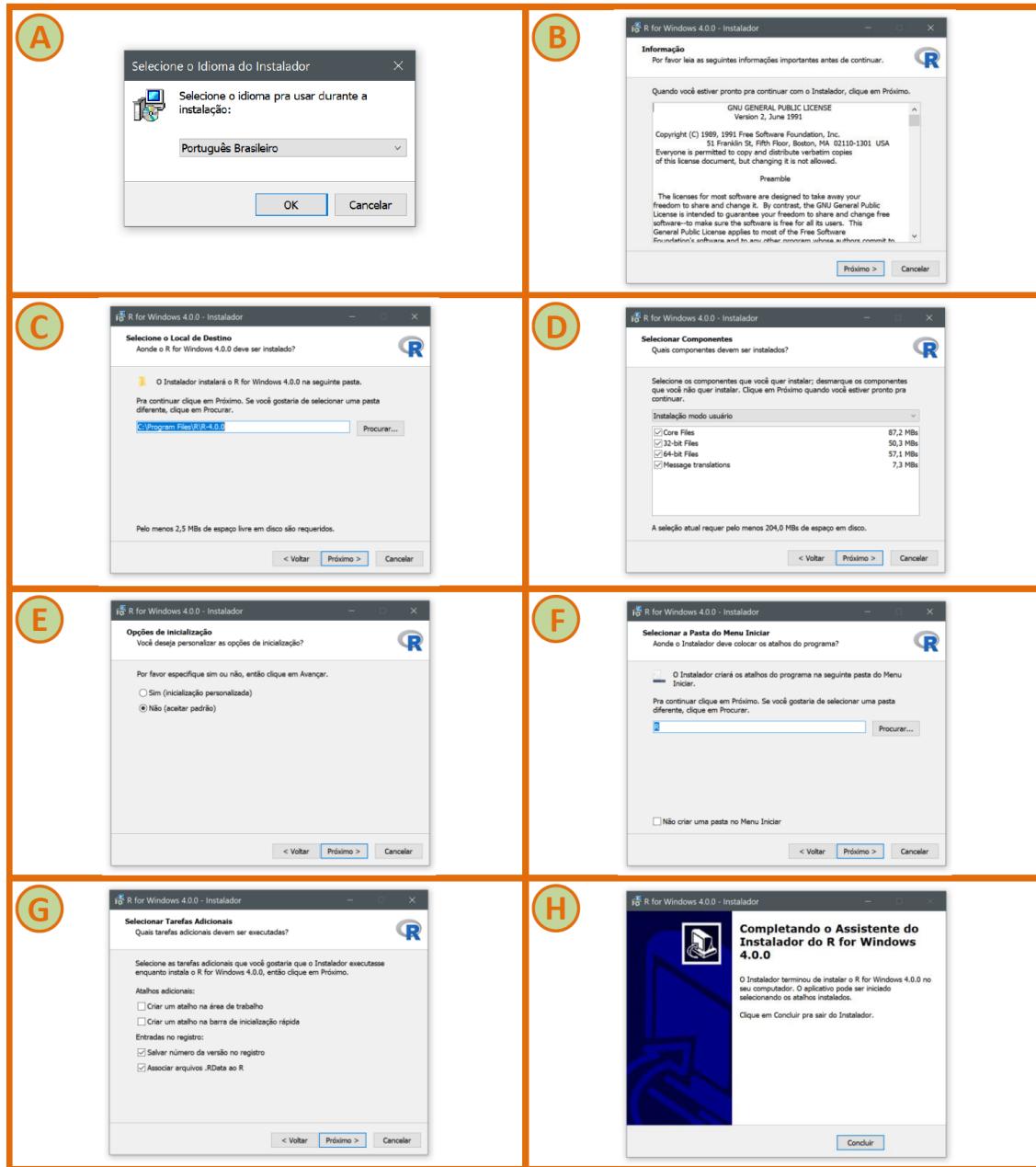


Figura 1.2: Passos para fazer a instalação do R

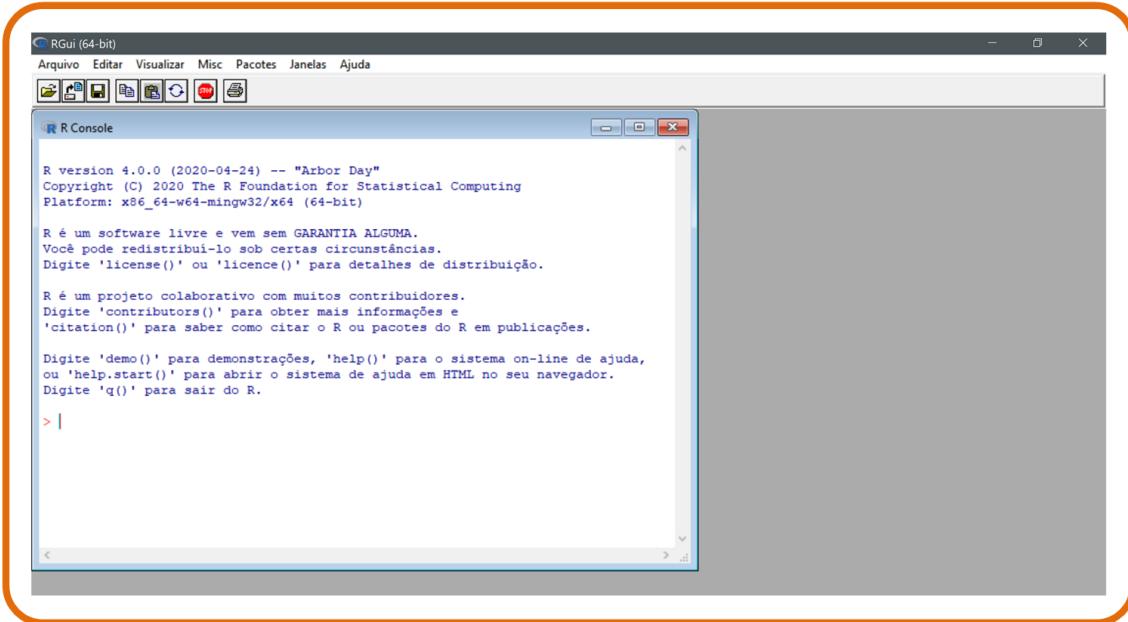


Figura 1.3: Aparência do software R



2. Download e instalação do RStudio

2.1 Sobre o RStudio

O RStudio é um ambiente de desenvolvimento integrado (IDE) para R. Ele inclui um console, editor de realce de sintaxe que suporta execução direta de código, além de ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho. Basicamente, o RStudio é uma interface gráfica mais prática e otimizada para o R. O software R é executado diretamente a partir do RStudio.

O RStudio está disponível em edições comerciais e de código aberto e é executado na área de trabalho (Windows, Mac e Linux) ou em um navegador conectado ao RStudio Server ou ao RStudio Server Pro (Debian / Ubuntu, Red Hat / CentOS e SUSE Linux). Maiores informações em <https://rstudio.com/about/>.

2.2 Download e instalação do RStudio

Neste material serão apresentados os passos para download e instalação do RStudio apenas para o sistema operacional Windows. Na Figura 2.1 podemos ver os passos para fazer o download do RStudio. Na página inicial do site do RStudio (Figuras 2.1 A e B) clique em "Download" e você será redirecionado para outra página (Figura 2.1 C). Nesta página desça a barra de rolagem localizada à direita até que apareçam as opções de download gratuitas e pagas (Figura 2.1 D) e clique na opção *RStudio Desktop Free* (Figura 2.1 E) para que você seja redirecionado para a página de download do RStudio Desktop 1.3.959 (Figuras 2.1 F e G). Nesta página clique em *DOWNLOAD RSTUDIO FOR WINDOWS* e será feito o download do arquivo de instalação do RStudio (Figura 2.1 H). Até o momento em que este material foi elaborado a versão mais recente do RStudio era a 1.3.959.

Após fazer o download do arquivo de instalação do RStudio (Figura 2.1 H) dê um clique duplo sobre ele para executá-lo. Nas janelas que irão abrir (Figuras 2.2 A à C) basta clicar em "Próximo" para fazer a instalação padrão do RStudio. Por fim clique em "Concluir" (Figura 2.2 D) para finalizar a instalação.

Após finalizar a instalação será criado um atalho para o RStudio na área de trabalho. Dê um



Figura 2.1: Passos para fazer o download do RStudio

clique duplo sobre o atalho para executar o RStudio.

Na Figura 2.3 pode-se observar a aparência do RStudio. Ao abrir o programa é mostrado um menu superior e quatro janelas. A janela superior esquerda é a janela do *Script*, a janela inferior esquerda é a janela do *Console*, a janela superior direita é responsável principalmente por listar os objetos gerados e a janela inferior direita é responsável principalmente por exibir os gráficos gerados. Caso não apareça a janela do *Script* na primeira seção do RStudio, você pode criar um script acessando o menu superior em **File** → **New File** → **R Script**.

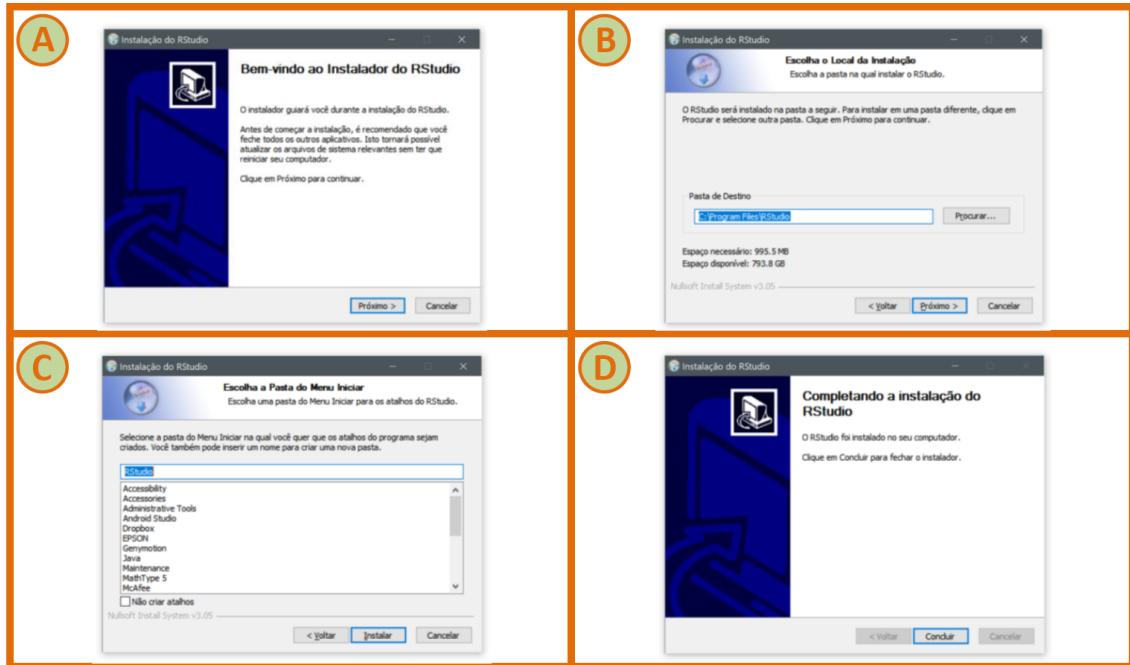


Figura 2.2: Passos para fazer a instalação do RStudio

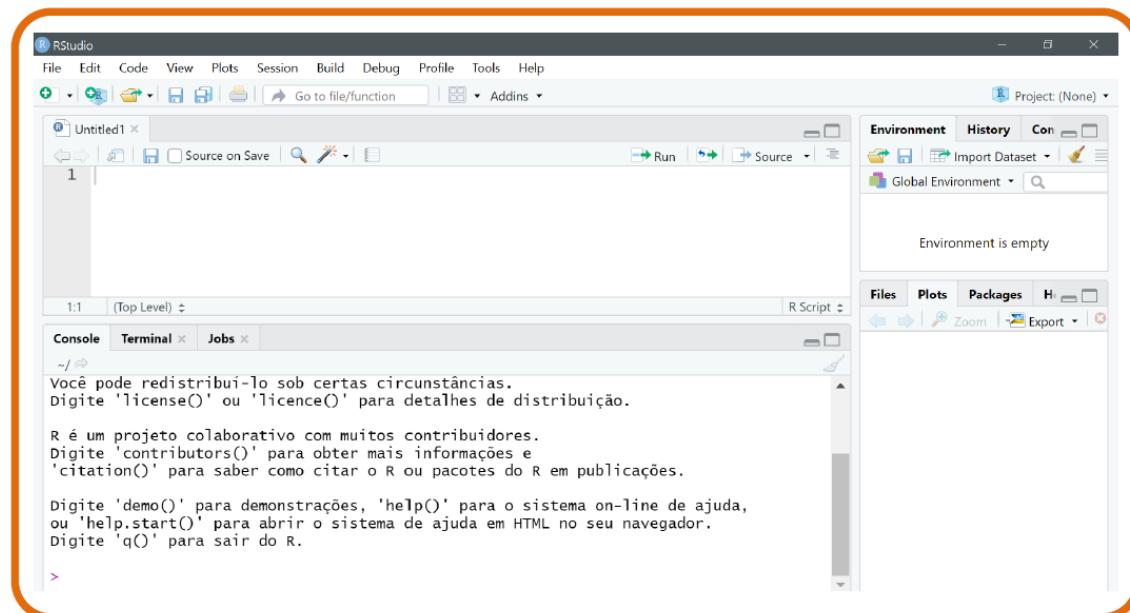


Figura 2.3: Aparência do RStudio



3. Introdução ao Software R

3.1 Console

A janela principal do R, que é iniciada junto com o programa é chamada de *R Console*. É nesta janela que os comandos digitados são executados. Para executar um comando no console basta digitar o comando e pressionar a tecla "Enter". Na Figura 3.1 a janela do console é a janela localizada no lado direito enquanto a janela localizada no lado esquerdo é chamada de *Script*.

3.2 Script

Uma desvantagem de se digitar os comandos diretamente no console é que depois de executar uma linha de comando não se pode editá-la nesta janela e fica difícil escrever e salvar todos os comandos de forma organizada.

Ao invés de digitar os comandos do R diretamente no console pode-se abrir uma nova janela no R chamada de *Script*. Nesta janela os comandos podem ser digitados e editados sem que eles sejam automaticamente executados, e, para que uma linha de comando do script seja enviada para o console e executada deve-se posicionar o cursor sobre a linha de comando do script e depois pressionar as teclas "Ctrl" e "R" juntas.

Na Figura 3.1 a janela localizada no lado esquerdo, inicialmente nomeada como "*Sem nome - Editor R*" (pelo fato de ainda não ter sido salva) é um script. Após salvar o script pode-se escolher um outro nome para ele.

Para abrir um novo script basta acessar o menu superior do R em: **Arquivo → Novo script**. Para salvar o script, contendo os comandos do R, a janela do script deve estar em uso (para isso basta clicar com o cursor sobre a janela do script) e depois deve-se acessar o menu superior do R em: **Arquivo → Salvar**.

Pode-se organizar as janelas do R (script e console) de forma que elas fiquem divididas lado à lado, como na Figura 3.1, ou ainda, divididas na horizontal. Para isso, basta acessar o menu superior do R em **Janelas → Dividir Lado a Lado** ou, ainda, **Janelas → Dividir na Horizontal**.

No RStudio também são usados scripts para digitar e editar os comandos. Você pode criar um

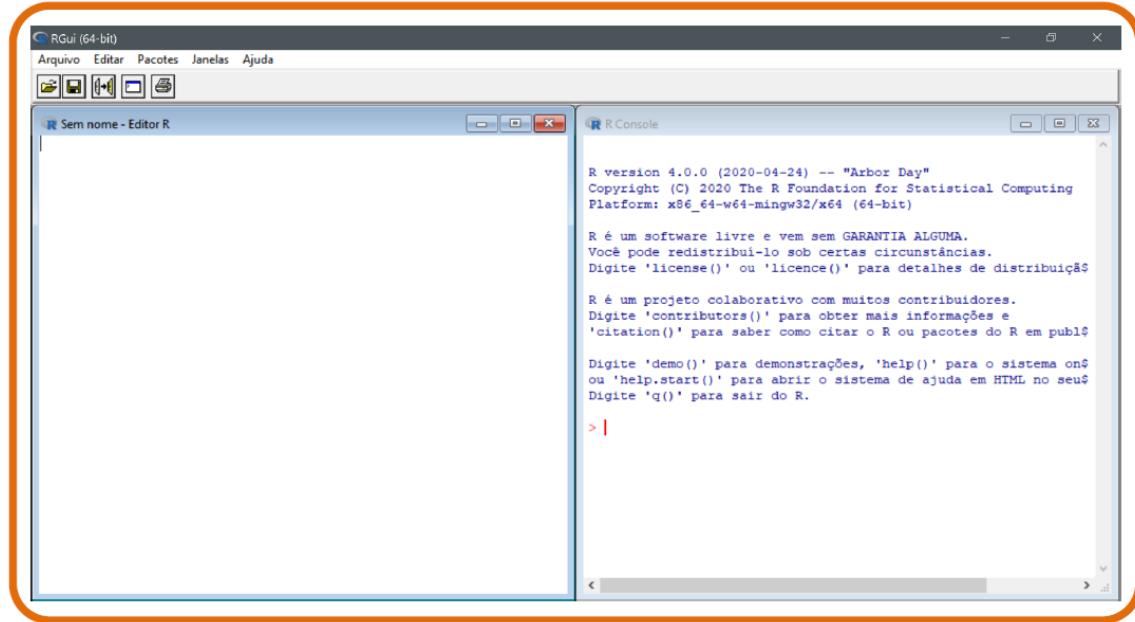


Figura 3.1: Software R com janelas divididas lado à lado

script no RStudio acessando o menu superior em **File -> New File -> R Script**. Para que uma linha de comando do script do RStudio seja enviada para o console e executada deve-se posicionar o cursor sobre a linha de comando do script e depois pressionar as teclas "Ctrl" e "Enter" juntas. Para salvar o script basta acessar o menu superior em **File -> Save As....**

3.3 Pacotes

Um elemento fundamental na estrutura do software R está relacionado com os comandos de execução de operações e técnicas, denominado de funções. As funções são disponibilizadas no R através de pacotes (*package*), também chamados de bibliotecas (*library*).

Os comandos básicos (funções básicas) já vem pré instalados e se encontram no pacote *base*. Há uma considerável oferta de pacotes com conteúdos diversos desenvolvidos por colaboradores. Para acessá-los é preciso realizar a instalação do pacote de interesse no R, pois diferente do pacote *base*, os demais pacotes não são disponibilizados na instalação do R.

A instalação de um pacote é realizada com o uso do comando `install.packages()`. Uma vez que o pacote é instalado deve-se fazer o seu carregamento na memória do programa, toda vez que se iniciar uma nova seção do R, com o uso do comando `library()`. Na Lista 3.1 são apresentados os comandos do software R para fazer a instalação e carregamento de um pacote (neste exemplo foi usado o pacote *agricolae*). Após o carregamento do pacote *agricolae*, todas as funções que compõem este pacote podem ser utilizadas pelo usuário.

```

1 # Instalação do pacote agricolae
2 install.packages("agricolae")
3
4 # Carregamento do pacote agricolae
5 library("agricolae")

```

Lista 3.1: Comandos do software R

3.4 Operações matemáticas básicas

Na Lista 3.2 são apresentados os comandos do software R para calcular alguns elementos de matemática básica.

```

1 # Adição
2 5+3
3
4 # Subtração
5 5-3
6
7 # Multiplicação
8 5*3
9
10 # Divisão
11 5/3
12
13 # Potenciação
14 5^3
15
16 # Raiz quadrada
17 sqrt(100)
18
19 # Exponencial
20 exp(0)
21 exp(1)
22
23 # Logarítmico na base e
24 log(1)
25 log(2)
26
27 # Logarítmico na base 10
28 log10(1)
29 log10(2)

```

Lista 3.2: Comandos do software R

3.5 Armazenamento de dados

Os tipos de dados mais comumente utilizados no R são os numéricos (dados compostos por números) e os caracteres (compostos por letras ou palavras). As informações imputadas no R são armazenadas na memória do programa e passam a ser denominadas *objetos*. Para criar um objeto basta associar um nome à informação de interesse, e, para isso, utiliza-se o símbolo `<-` ou `=` entre o nome do objeto e a informação que define o objeto. No caso de objeto formado por caracteres a informação armazenada deve vir entre aspas.

Na Lista 3.3 são apresentados os comandos do software R para fazer a entrada de dados numéricos e de caracteres. Observe, no último comando da lista, que pode-se usar o ";" para colocar dois comandos diferentes (criar o objeto e mostrar o objeto) em uma mesma linha.

```

1 # Entrando com um dado numérico (criando o objeto x)
2 x <- 10
3
4 # Mostrando o objeto x
5 x
6
7 # Entrando com um dado de caracteres (criando o objeto y)

```

```

8 # e mostrando o objeto (comandos na mesma linha separados por ";")
9 y <- "bola" ; y

```

Lista 3.3: Comandos do software R

Algumas regras devem ser seguidas na hora de nomear um objeto. O nome do objeto precisa começar com uma letra, não pode ser um número, não pode conter símbolos referentes a funções ou nome de funções e a "seta" `<-` deve estar sempre apontada para o nome do objeto.

Observação: O R diferencia letras maiúsculas e minúsculas. Assim, por exemplo, ao criar um objeto "x" (minúsculo) deve-se sempre se referir a ele com a letra x minúscula pois, se for utilizada a letra X (maiúscula) o R irá retornar uma mensagem de erro: "Erro: objeto 'X' não encontrado".

3.6 Operadores lógicos

Na Tabela 3.1 são apresentados alguns operadores lógicos que são usados para realização de "testes" entre dois objetos.

Tabela 3.1: Operadores lógico

Operador	Descrição
<code>==</code>	igual a
<code>!=</code>	diferente de
<code><</code>	menor que
<code>></code>	maior que
<code><=</code>	menor ou igual a
<code>>=</code>	maior ou igual a
<code>x y</code>	x OU y
<code>x & y</code>	x E y

Na Lista 3.4 são apresentados alguns exemplos de comandos do software R utilizando os operadores lógicos apresentados na Tabela 3.1.

```

1 # Criando objetos x e y
2 x <- c(1,3,2,4,5,3,2,4); x
3 y <- c(1,2,3,4,5,6,7,8); y
4
5 # Operador "igual a"
6 x==3
7
8 # Operador "diferente de"
9 x!=3
10
11 # Operador "maior que"
12 x>3
13
14 # Operador "maior ou igual a"
15 x>=3
16
17 # Operador "OU"
18 if(x[2] | y[2] > 2) print("sim") else print("não")

```

```

19 # Operador "E"
20 if(x[2] & y[2] > 2) print("sim") else print("não")
21

```

Lista 3.4: Comandos do software R

Ao executar os comandos das linhas 6, 9, 12 e 15 da Lista 3.4 o R irá verificar quais elementos do objeto `x` satisfazem as condição impostas e retornará TRUE se o elemento satisfaz a condição e FALSE se não satisfaz. O comando da linha 18 irá verificar se o segundo elemento do objeto `x` **OU** o segundo elemento do objeto `y` é maior do que o número 2, se sim retornará "sim" e se não retornará "não", ou seja, se um elemento ou outro for maior que 2 o R retornará "sim". Já o comando da linha 21 irá verificar se o segundo elemento do objeto `x` **E** o segundo elemento do objeto `y` são maiores do que o número 2, se sim retornará "sim" e se não retornará "não", ou seja, apenas se os dois elementos forem maiores que 2 o R retornará "sim".

3.7 Arquivos do R

O R utiliza um diretório (pasta) do computador para gravar, ler, importar e exportar arquivos. Para saber o diretório em uso digite o comando: `getwd()`. Para alterar o diretório de trabalho digite: `setwd("C:/Endereço")`. Os principais tipos específicos de arquivos do R são: o `*.RData` e o `*.Rhistory`.

O arquivo com extensão `*.RData` representa a área de trabalho na qual são salvos os objetos criados durante uma seção do R. Para salvar uma área de trabalho basta acessar o menu superior do R em: **Arquivo → Salvar área de trabalho...**

O arquivo com extensão `*.Rhistory`, denominado de histórico, armazena todos os comandos utilizados em uma seção R. Os arquivos de histórico apresentam o formato de texto e por isso podem ser visualizados como um arquivo `*.txt`. Para salvar um histórico basta acessar o menu superior do R em: **Arquivo → Salvar Histórico...**

Um Script é uma janela onde os comandos do R podem ser digitados e editados sem que sejam executados. Para abrir um novo script basta acessar o menu superior do R em: **Arquivo → Novo script**. Para salvar o script contendo os comandos do R basta clicar sobre a janela do script e depois acessar o menu superior do R em: **Arquivo → Salvar**. Os scripts apresentam o formato de texto e por isso podem ser visualizados como um arquivo `*.txt`. Caso o R esteja em uso a partir do RStudio basta salvar o script para que os comandos utilizados sejam salvos.

3.8 Estrutura de dados

No R, os dados contidos em um objeto podem assumir distintas estruturas. As mais utilizadas são os vetores, as matrizes e os data-frames.

3.8.1 Vetores

Um vetor é a forma mais simples de armazenamento de dados. Pode ser um vetor numérico ou de caracteres. A função `c()` é utilizada na criação do vetor. Na Lista 3.5 são apresentados os comandos do software R para criar um vetor numérico e um vetor de caracteres.

```

1 # Vetor com dados numéricos
2 x <- c(10, 20, 30)
3

```

```

4 # Mostrando o vetor x
5 x
6
7 # Vetor de caracteres
8 y <- c("João", "Ana", "Pedro Henrique")
9
10 # Mostrando o vetor y
11 y

```

Lista 3.5: Comandos do software R

No caso de um vetor de caracteres representar uma variável categórica (qualitativa) é possível associar ao vetor os atributos de fator e os níveis (categorias) do fator usando as funções `factor()` e `levels()`. Na Lista 3.6 são apresentados os comandos do software R para converter um vetor de caracteres em fator e mostrar os níveis do fator.

```

1 # Convertendo um vetor de caracteres em fator
2 w <- factor(c("Masculino", "Feminino", "Masculino"))
3
4 # Mostrando os níveis
5 levels (w)

```

Lista 3.6: Comandos do software R

Na Lista 3.7 são apresentados os comandos do software R para ordenar os níveis de um fator. Isto será útil mais adiante.

```

1 # Criando um vetor convertido em fator com níveis não ordenados
2 cs <- factor(c("media", "baixa", "media", "alta", "baixa", "baixa", "alta",
   "media", "alta", "media"))
3
4 # Ordenando os níveis do fator cs
5 cs <- factor(cs, ord=T, levels=c("baixa", "media", "alta"))
6
7 # Mostrando o objeto cs
8 cs

```

Lista 3.7: Comandos do software R

Manipulação de vetores

A função `seq()` cria um vetor com uma sequência de números em um intervalo especificado. Na Lista 3.8 são apresentados os comandos do software R para criar vetores de números usando a função `seq()`.

```

1 # Sequência de 1 a 50, com intervalo igual a 1
2 seq(1, 50, 1)
3
4 # Sequência de 1 a 50, com intervalo igual a 5
5 seq(1, 50, 5)

```

Lista 3.8: Comandos do software R

A função `rep()` cria um vetor com uma números repetidos. Na Lista 3.9 são apresentados os comandos do software R para criar vetores de números usando a função `rep()`.

```

1 # Repetição do número 1 três vezes
2 rep(1, 3)
3
4 # Repetição dos números de 1 a 5 duas vezes
5 rep(1:5, 2)
6
7 # Repetição dos números 0 e 1, repetidos 10 vezes, alternadamente
8 rep(c(0, 1), 10)
9
10 # Repetição dos números 0 e 1, repetidos 10 vezes, sem alternar
11 rep(c(0, 1), each=10)
12
13 # ou ainda
14 c(rep(0, 10), rep(1, 10))
15
16 # Repetição da palavra "Sim" dez vezes
17 rep("Sim", 10)

```

Lista 3.9: Comandos do software R

Seleção de elementos dentro de um vetor. Para isso são utilizados colchetes com a indicação das posições ou os próprios elementos. Na Lista 3.10 são apresentados os comandos do software R para selecionar elementos de vetores.

```

1 # Seleciona o primeiro elemento do vetor x
2 x <- c(10, 20, 30, 40, 50)
3 x[1]
4
5 # Seleciona os elementos do vetor x nas posições 1, 3, e 5
6 x[c(1, 3, 5)]
7
8 # Seleciona os elementos do vetor x nas posições 2 e 4
9 x[c(2, 4)]
10
11 # Seleciona os elementos do vetor x que não estão nas posições 2 e 4
12 x[-c(2, 4)]
13
14 # Seleciona os elementos do vetor x que forem iguais a 20
15 x[x==20]
16
17 # Seleciona o mínimo entre os elementos do vetor x
18 x[x==min(x)]
19
20 # Seleciona os elementos do vetor x maiores do que 20
21 x[x>20]
22
23 # Verifica quais elementos são maior que 20 (TRUE é porque o elemento
24     # satisfaz a condição ">20" e FALSE não satisfaz)
25 x>20
26
27 # Mostra o comprimento de um vetor (número de elementos do vetor)
length(x)

```

Lista 3.10: Comandos do software R

Operações com vetores

Existe a possibilidade de somar, subtrair, dividir ou multiplicar vetores numéricos. Cada valor de uma posição de um vetor é somado, subtraído, dividido ou multiplicado pelo valor da posição

correspondente do outro vetor. Também é possível Na Lista 3.11 são apresentados os comandos do software R para fazer operações com vetores.

```

1 # Criando dois vetores x e y
2 x <- c(1, 2, 3, 4, 5)
3 y <- c(3, 8, 4, 7, 2)
4
5 # Mostrando os vetores x e y
6 x; y
7
8 # Somando x e y
9 x+y
10
11 # Multiplicando x e y (multiplica cada elemento do vetor x pelo elemento da
12 # respectiva posição no vetor y)
12 x*y
13
14 # Dividindo x por y (divide cada elemento do vetor x pelo elemento da
14 # respectiva posição no vetor y)
15 x/y
16
17 # Somando o número 10 a todos os elementos do vetor x
18 10+x
19
20 # subtraindo o número 10 de todos os elementos do vetor x
21 x-10
22
23 # Multiplicando o vetor x por 10
24 10*x
25
26 # Dividindo o vetor x por 10
27 x/10

```

Lista 3.11: Comandos do software R

3.8.2 Matrizes

No R uma Matriz é montada a reorganizando os elementos de um vetor em linhas e colunas. Por padrão o R monta a matriz por colunas. Para inverter este padrão, ou seja, para preencher uma matriz por linhas deve-se usar o argumento `byrow=T`. Também podemos criar uma matriz a partir de um ou mais vetores. Na Lista 3.12 são apresentados os comandos do software R para construir matrizes e acessar elementos da matriz.

```

1 # Matriz criada a partir do vetor que varia de 1 a 9. A matriz é composta
# pelo preenchimento sucessivo das colunas
2 matrix(1:9, nrow=3)
3
4 # A matriz é composta pelo preenchimento sucessivo das linhas
5 matrix(1:9, nrow=3, byrow=T)
6
7 # Matriz formada por dois vetores, organizada por linhas
8 x <- c(1, 2, 3)
9 y <- c(10, 20, 30)
10 rbind(x, y)
11
12 # Matriz formada por dois vetores, organizada por colunas

```

```

13 cbind(x, y)
14
15 # Criando o objeto X como uma matriz
16 X <- matrix(1:9, ncol=3)
17
18 # Mostrando a matriz X
19 X
20
21 # Acessando a primeira coluna da matriz X
22 X[,1]
23
24 # Acessando a segunda linha da matriz X
25 X[2,]
26
27 # Acessando o elemento X23
28 X[2,3]

```

Lista 3.12: Comandos do software R

Operações com matrizes

Na Lista 3.13 são apresentados os comandos do software R para fazer operações com matrizes.

```

1 # Criando a matriz X
2 X <- matrix(1:9, ncol=3); X
3
4 # Criando a matriz y
5 Y <- matrix(rep(1:3, 3), ncol=3); Y
6
7 # Transposta de X
8 Xt <- t(X); Xt
9
10 # Multiplicação de elemento por elemento
11 X*Y
12
13 # Multiplicação entre matrizes
14 X%*%Y

```

Lista 3.13: Comandos do software R

3.8.3 Data-frame

O *data-frame* é uma estrutura de dados semelhante a uma matriz, com a diferença que pode armazenar caracteres além de números.

As informações provenientes de uma pesquisa, geralmente, seguem a estrutura de um data-frame no qual colunas representam as variáveis da pesquisa e as linhas os indivíduos/unidades experimentais.

Na Lista 3.14 são apresentados comandos do software R para criar um exemplo de data-frame.

```

1 # Criando vetores de dados numéricos e de caracteres
2 nome <- c("Júlia", "João", "Isabela", "Gustavo")
3 idade <- c(20, 18, 22, 21)
4 peso <- c(50, 63, 60, 80)
5 sexo <- factor(c("F", "M", "F", "M"))
6
7 # Criando o data.frame

```

```

8   dados <- data.frame(nome, idade, peso, sexo)
9
10  # Mostrando o data.frame
11  dados

```

Lista 3.14: Comandos do software R

Na Lista 3.15 são apresentados comandos do software R para manipular o data.frame gerado na Lista 3.14.

```

1  # Acessar a coluna idade do data.frame pelo nome da coluna
2  dados$idade
3
4  # Selecionar indivíduos com idade superior a 20 anos
5  dados1 <- dados[dados$idade>20,];dados1
6
7  # Selecionar indivíduos do sexo feminino
8  dados[dados$sexo=="F",]
9
10 # Selecionar indivíduos com peso diferente de 60
11 dados[dados$peso!=60,]
12
13 # Selecionar dados apenas de Júlia
14 dados[dados$nome=="Júlia",]
15 dados[1,] # outra forma
16
17 # Ordenar o data.frame, em ordem crescente, de acordo com a coluna idade
18 dados[order(dados$idade),]
19
20 # Ordenar o data.frame, em ordem decrescente, de acordo com a coluna idade
21 dados[rev(order(dados$idade)),]

```

Lista 3.15: Comandos do software R

3.8.4 Listas

Listas são estruturas genéricas que permitem armazenar diversos tipos de estruturas de dados em um único objeto.

Na Lista 3.16 são apresentados comandos do software R para gerar um exemplo de objeto do tipo lista.

```

1  # Gerando um objeto do tipo lista
2  lista1 <- list(Vetor = 1:10, Caracter = "Análise Exploratória de Dados",
3                 Matriz = matrix(1:16, 4, 4))
4
5  # Mostrando o objeto
6  lista1
7
8  # Acessando cada estrutura da lista individualmente
9  lista1$Vetor
10  lista1$Caracter
11  lista1$Matriz
12
13  # Outra forma de acessar cada estrutura
14  lista1[1]
15  lista1[2]
16  lista1[3]

```

Lista 3.16: Comandos do software R

3.9 Importação de conjunto de dados para o R

O R possui uma série de comandos para a importação de banco de dados. Os comandos de leitura de banco de dados seguem a seguinte estrutura:

```
função("local do arquivo/nome do arquivo", argumentos do comando)
```

Argumentos:

- dec = " " -> especifica o separador decimal, que pode ser ". " ou ", ".
- sep = " " -> especifica o separador da coluna, que pode ser ";" ou ", ".
- header = TRUE ou h = T -> especifica que a primeira linha é cabeçalho (contém os nomes das colunas), caso contrário use FALSE.

Na Lista 3.17 são apresentados comandos do software R para importar alguns tipos de arquivos de dados. Os arquivos de dados utilizados podem ser baixados em <https://drive.google.com/drive/folders/1tCzWnInYGOdpC0qQsJ2EvJLDqLQ7z8aB?usp=sharing>.

```
1 # Formato .txt
2 # Separador de colunas: tabulação
3 # Separador decimal: dec = "." (ponto)
4 # Primeira linha é cabeçalho: header = TRUE
5 dados1 <- read.table("C:/caminho_do_diretorio/data1.txt", header=TRUE)
6 dados1
7
8 # Formato .txt
9 # Separador de colunas: sep = "," (vírgula)
10 # Separador decimal: dec = "." (ponto)
11 # Primeira linha é cabeçalho: header = TRUE
12 dados2 <- read.table("C:/caminho_do_diretorio/data2.txt",
13     sep=",", dec=". ", h=T)
14 dados2
15
16 # Formato .csv
17 # Separador de colunas: sep = "," (vírgula)
18 # Separador decimal: dec = "." (ponto)
19 # Primeira linha é cabeçalho: header = TRUE
20 dados3 <- read.csv("C:/caminho_do_diretorio/data3.csv", h=T)
21 dados3
22
23 # Formato .csv
24 # Separador de colunas: sep = ";" (ponto e vírgula)
25 # Separador decimal: dec = "," (vírgula)
26 # Primeira linha é cabeçalho: header = TRUE
27 dados4 <- read.csv2("C:/caminho_do_diretorio/data4.csv", h=T)
28 dados4
29
30 # Formato .xlsx (Excel)
31
32 # Carregando o pacote "readxl" (tem que instalar este pacote antes)
33 library("readxl")
34
35 # Separador decimal: dec = "," (vírgula)
36 # Primeira linha é cabeçalho
37 dados5 <- read_xlsx("C:/caminho_do_diretorio/data5.xlsx")
38 dados5
39
40
```

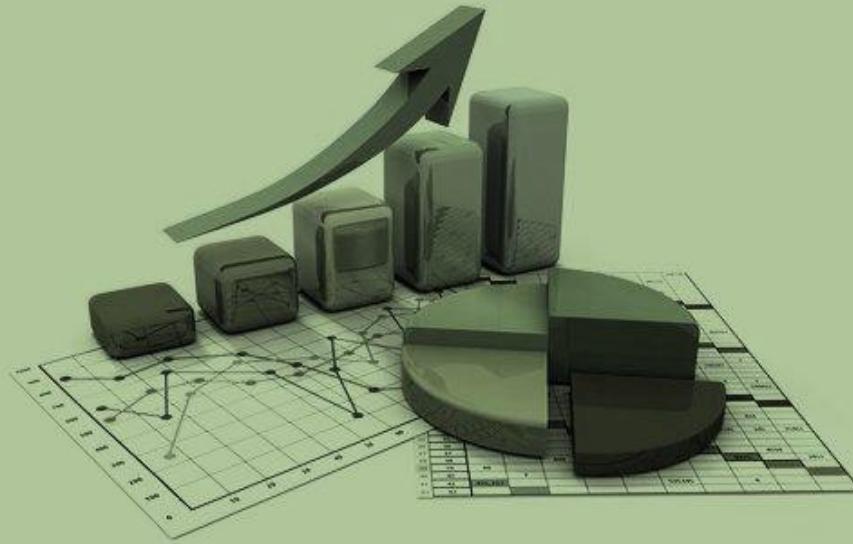
```
41 # Mudando o diretório atual do R  
42 setwd("C:/caminho_do_diretorio")  
43  
44 # Verificando o diretório atual do R  
45 getwd()  
46  
47 # Verificando os arquivos no diretório atual do R  
48 dir()  
49  
50 # Importando dados diretamente do diretório atual do R  
51 # Nesse caso não precisa especificar o caminho do diretório  
52 dados6 <- read.table("data1.txt", header = TRUE)  
53 dados6
```

Lista 3.17: Comandos do software R

Observação: Na Lista 3.17 deve-se substituir "C:/caminho_do_diretorio/" pelo caminho até o diretório (pasta) do computador em que foram salvos os dados. Outra observação importante é que ao digitar o caminho do diretório deve-se usar barras à direita: "/".

Seção II

4	Tipos de variáveis	31
4.1	Variáveis quantitativas	
4.2	Variáveis qualitativas	
5	Tabelas	33
5.1	Normas para construção de tabelas	
5.2	Tabelas de distribuição de frequências	
6	Gráficos	43
6.1	Gráficos de colunas e barras	
6.2	Gráfico de setores	
6.3	Diagrama de linhas	
6.4	Histograma	
6.5	Polygono de frequências	
6.6	Ramo e folhas	
6.7	Gráfico de Série Temporal	
7	Medidas de posição	61
7.1	Introdução	
7.2	Somatório	
7.3	Média aritmética	
7.4	Mediana	
7.5	Moda	
7.6	Utilização das medidas de tendência central	
7.7	Medidas separatrizes (quantis)	
7.8	Boxplot	
8	Medidas de dispersão	89
8.1	Amplitude	
8.2	Desvio médio absoluto	
8.3	Variância	
8.4	Desvio padrão	
8.5	Coeficiente de variação	
8.6	Propriedades das Medidas de Dispersão	
9	Assimetria e curtose	101
9.1	Assimetria	
9.2	Curtose	
10	Análise Bivariada	109
10.1	Variáveis qualitativas	
10.2	Variáveis quantitativas	
	Bibliografia	123



4. Tipos de variáveis

Uma característica que pode assumir diferentes valores de um indivíduo para outro é chamada de variável. Por exemplo, a característica altura é uma variável pois diferentes indivíduos podem apresentar diferentes alturas. As variáveis podem ser classificadas em qualitativas e quantitativas. Ainda, as variáveis qualitativas podem ser classificadas em nominais e ordinais, já as variáveis quantitativas podem ser classificadas em discretas e contínuas. Pode-se observar na Figura 4.1 a classificação das variáveis.

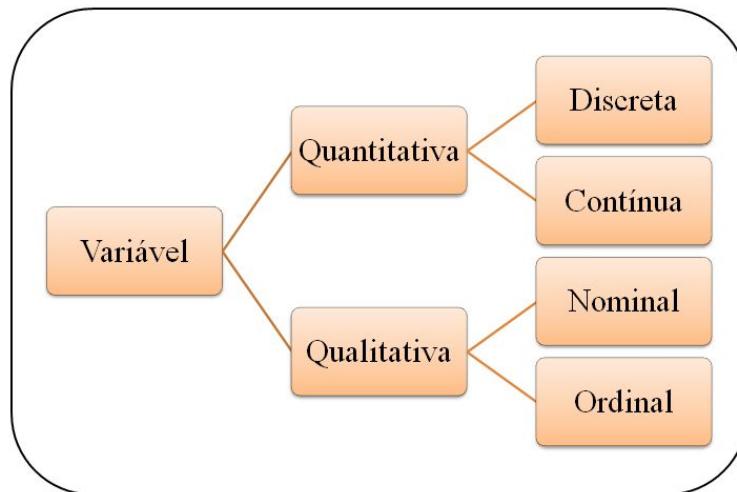


Figura 4.1: Classificação das variáveis

4.1 Variáveis quantitativas

São variáveis cujas possíveis realizações são números resultantes de uma contagem ou mensuração.

- **Variáveis Quantitativas Discretas:** São variáveis numéricas para as quais os possíveis valores formam um conjunto finito ou enumerável de números, e que resultam, frequentemente, de uma contagem. Por exemplo, a variável número de filhos, cujas possíveis realizações são 0, 1, 2, 3, ..., é uma variável quantitativa discreta.
- **Variáveis Quantitativas Contínuas:** São variáveis numéricas para as quais os possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração. Por exemplo, a variável altura de um indivíduo, cujas possíveis realizações são números reais positivos (por exemplo: 1,60m, 1,56m, 1,75m, ...) é uma variável quantitativa contínua.

4.2 Variáveis qualitativas

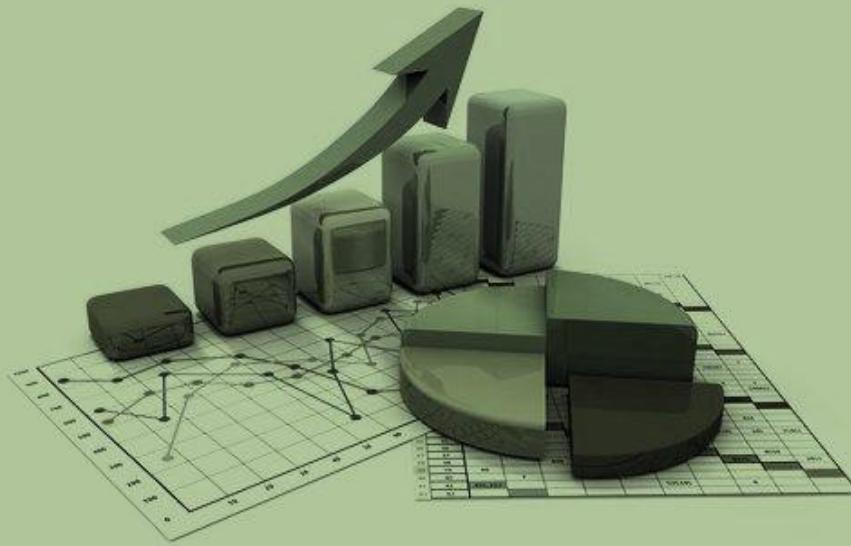
São variáveis que apresentam como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado.

- **Variáveis qualitativas nominais:** São variáveis cujas possíveis realizações são atributos para os quais não existe nenhuma ordenação. Por exemplo, a variável sexo, cujas possíveis realizações são masculino e feminino, é uma variável qualitativa nominal pois suas realizações não tem nenhuma ordenação.
- **Variáveis qualitativas ordinais:** São variáveis cujas possíveis realizações são atributos para os quais existe uma ordem. Por exemplo, a variável grau de instrução, cujas possíveis realizações podem ser: ensino fundamental, ensino médio e ensino superior; é uma variável qualitativa ordinal pois suas possíveis realizações seguem uma ordem crescente que está atrelada aos anos de estudo concluídos.

■ **Exemplo 4.1** Classifique as variáveis apresentadas a seguir:

- | | |
|---------------------|---------------------------------|
| a) Peso | Resposta: quantitativa contínua |
| b) Classe social | Resposta: qualitativa ordinal |
| c) Número de irmãos | Resposta: quantitativa discreta |
| d) Tempo | Resposta: quantitativa contínua |
| e) Cor dos olhos | Resposta: qualitativa nominal |
| f) Raça de gatos | Resposta: qualitativa nominal |

■



5. Tabelas

5.1 Normas para construção de tabelas

Os dados são apresentados em tabelas colocadas perto do ponto do texto em que são mencionadas pela primeira vez. As tabelas devem conter os seguintes elementos: título, cabeçalho, indicador de linha, célula e moldura, como mostrado no exemplo a seguir.

■ **Exemplo 5.1** Considerando a Tabela 5.1, apresentada a seguir, serão ilustrados os elementos que compõem uma tabela.

Tabela 5.1: População do Brasil, segundo o sexo, de acordo com o Censo Demográfico 2010

Sexo	População residente
Homens	93.406.990
Mulheres	97.348.809
Total	190.755.799

Fonte: Censo Demográfico 2010. IBGE (2011).

Elementos de uma tabela

- **Título:** especifica o conteúdo da tabela.

Tabela 5.1: População do Brasil, segundo o sexo, de acordo com o Censo Demográfico 2010

O título deve ser colocado na parte superior da tabela.

- **Cabeçalho:** especifica o conteúdo de cada coluna.

Sexo	População residente
------	---------------------

- **Indicador de linha:** é um conjunto de termos em que cada termo indica o conteúdo de uma linha.

Homens
Mulheres
Total

- **Célula:** resulta do cruzamento de uma linha com uma coluna e deve conter um dado numérico.

93.406.990
97.348.809
190.755.799

Nenhuma célula da tabela deve ficar em branco. Toda célula deve apresentar um número ou, se o dado não existir, deve-se colocar um traço (–) na célula para indicar que aquele dado não existe.

- **Moldura:** Entende-se por moldura o conjunto de traços que delimitam a tabela.



A moldura é composta apenas por traços horizontais. Basicamente é composta por um traço superior e um inferior que delimitam a tabela e mais um traço abaixo do cabeçalho para delimitá-lo. Quando a última linha da tabela representa a soma das colunas (linha do Total) é costume fazer mais um traço horizontal acima desta linha para delimitá-la também. Pode-se, ainda, fazer traços verticais no interior da tabela para separar uma coluna da outra, porém, não se deve "fechar" as laterais da tabela com traços verticais.

- **Fonte:** indica o responsável (pessoa física ou jurídica) pelos dados.

Fonte: Censo Demográfico 2010. IBGE (2011).

A Fonte deve ser colocada na primeira linha do rodapé da tabela e precedida pela palavra *Fonte*. Não é necessário indicar a fonte nos casos em que os dados forem obtidos pelo próprio pesquisador.

5.2 Tabelas de distribuição de frequências

5.2.1 Distribuição de frequências para dados qualitativos

Quando observamos dados qualitativos, classificamos cada observação em determinada categoria. Depois, contamos o número de observações em cada categoria. A idéia é resumir as informações na forma de uma tabela que mostre essas contagens (frequências) por categoria.

■ **Exemplo 5.2** Considere os dados brutos do grau de instrução de 36 indivíduos:

Fundamental	Médio	Médio	Superior	Médio	Médio
Médio	Médio	Fundamental	Fundamental	Superior	Fundamental
Superior	Fundamental	Médio	Médio	Médio	Médio
Fundamental	Médio	Superior	Fundamental	Superior	Médio
Médio	Fundamental	Médio	Médio	Médio	Fundamental
Fundamental	Fundamental	Superior	Fundamental	Médio	Médio

Contando os dados por categoria temos 12 indivíduos cujo grau de instrução é o ensino fundamental, 18 indivíduos com ensino médio e 6 indivíduos com ensino superior. Assim, podemos resumir essas informações na Tabela 5.2.

Tabela 5.2: Distribuição de frequências do grau de instrução de 36 indivíduos

Grau de instrução	Frequências (f_i)
Fundamental	12
Médio	18
Superior	6
Total	36

As tabelas de distribuição de frequências podem conter, além das frequências (também conhecidas como frequências absolutas: f_i), as proporções (também conhecidas como frequências relativas: fr_i) e as porcentagens (também conhecidas como frequências percentuais: fp_i). As frequências relativas são calculadas dividindo-se cada frequência absoluta pelo total, enquanto as frequências percentuais são calculadas multiplicando-se cada frequência relativa por 100. Por exemplo, para o grau de instrução ensino fundamental, obtemos a frequência relativa fazendo $12/36 = 0,3333$, e a frequência percentual fazendo $0,3333 \times 100 = 33,33\%$, como pode-se observar na Tabela 5.3.

Tabela 5.3: Distribuição de frequências do grau de instrução de 36 indivíduos

Grau de instrução	f_i	fr_i	fp_i
Fundamental	12	0,3333	33,33%
Médio	18	0,5000	50,00%
Superior	6	0,1667	16,67%
Total	36	1,0000	100,00%

Na Lista 5.1 são apresentados os comandos do software R para organizar os dados do Exemplo 5.2 em uma distribuição de frequências.

```

1 # Entrando com os dados no R:
2 dados <- c("Fundamental", "Médio", "Médio", "Superior", "Médio", "Médio",
3         "Médio", "Médio", "Fundamental", "Fundamental", "Superior", "Fundamental",
4         "Superior", "Fundamental", "Médio", "Médio", "Médio", "Médio",
5         "Fundamental", "Médio", "Superior", "Fundamental", "Superior", "Médio",
6         "Médio", "Fundamental", "Médio", "Médio", "Fundamental",
7         "Fundamental", "Fundamental", "Superior", "Fundamental", "Médio")
8
9 # Mostrando os dados armazenados (OPCIONAL)
10 dados
11
12 # Gerando a distribuição de frequências
13 tab <- table(dados)
14
15 # Mostrando a distribuição de frequências
16 tab
17
18 # Gerando e mostrando as proporções
19 tabprop <- prop.table(tab) ; tabprop

```

Lista 5.1: Comandos do software R

Observação: ao executar os comandos da Lista 5.1, os níveis da variável analisada aparecem na tabela de distribuição de frequências em ordem alfabética: "Fundamental", "Médio", "Superior". Neste caso, a ordem alfabética também coincide com a ordem natural dos níveis dessa variável. Porém, se a ordem dos níveis de uma variável qualitativa ordinal não coincidir com a ordem alfabética, no R também é possível ordenar as categorias desta variável ao gerar uma tabela de distribuição de frequências.

Na Lista 5.2 são apresentados os comandos para ordenar os níveis da variável (fator) "classe social" e assim gerar uma tabela de distribuição de frequências com os níveis desse fator devidamente ordenados.

```

1 # Criando um vetor convertido em fator com níveis não ordenados
2 cs <- factor(c("media", "baixa", "media", "alta", "baixa", "baixa", "alta",
3               "media", "alta", "media"))
4
5 # Gerando uma distribuição de frequências sem ordenar os níveis desse fator
6 table(cs)
7
8 # Ordenando os níveis do fator cs (classe social)
9 cs <- factor(cs, ord=T, levels=c("baixa", "media", "alta"))
10
11 # Gerando uma distribuição de frequências com os níveis do fator ordenados
12 table(cs)

```

Lista 5.2: Comandos do software R

Observação: ao executar os comandos da Lista 5.2, na saída da linha 5 os níveis do fator "grau" apresentados na tabela de distribuição de frequências aparecem na ordem: "alto", "baixo", "medio". Já na saída da linha 11 eles aparecem na ordem: "baixo", "medio", "alto", pois os níveis do fator foram previamente ordenados.

5.2.2 Distribuição de frequências para dados quantitativos

Dados quantitativos também podem ser agrupados em tabelas de distribuição de frequências, porém a construção da tabela é diferente para dados discretos e contínuos.

Tabela para dados quantitativos discretos

Para organizar uma tabela de distribuição de frequências para dados quantitativos discretos deve-se seguir os passos:

1. escreva os dados em ordem crescente (rol);
2. conte quantas vezes cada valor se repete (frequência);
3. organize a tabela apresentando os valores numéricos em ordem natural e suas respectivas frequências.

■ **Exemplo 5.3** Considere os dados brutos do número de faltas ao trabalho de trinta funcionários:

1	3	1	1	0	1	0	1	1	0	2	2	0	0	0
1	2	1	2	0	0	1	6	4	3	3	1	2	4	0

Construa uma tabela de distribuição de frequências do número de faltas ao trabalho.

Solução

Organizando os dados em ordem crescente temos:

0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
1	1	1	1	2	2	2	2	2	3	3	3	4	4	6

Agora basta contar as frequências (quantidade de vezes que cada número ocorre) e montar a tabela de distribuição de frequências (Tabela 5.4).

Observação: Como o número de faltas $x_i = 5$ não aparece no conjunto de dados brutos ele deve ser representado na tabela de distribuição de frequências como tendo frequência $f_i = 0$.

Tabela 5.4: Distribuição de frequências do número de faltas ao trabalho de trinta funcionários

Número de faltas	f_i
0	9
1	10
2	5
3	3
4	2
5	0
6	1
Total	30

Na Lista 5.3 são apresentados os comandos do software R para organizar os dados do Exemplo 5.3 em uma distribuição de frequências.

```

1 # Entrando com os dados no R
2 dados <- c(1,3,1,1,0,1,0,1,1,0,2,2,0,0,0,1,2,1,2,0,0,1,6,4,3,3,1,2,4,0)
3
4 # Mostrando os dados armazenados
5 dados
6
7 # Tabela de distribuição de frequências
8 tab <- table(dados)
9
10 # Mostrando a tabela de distribuição de frequências
11 tab

```

Lista 5.3: Comandos do software R

Observação: Ao executar os comandos do software R apresentados na Lista 5.3 não irá aparecer na saída do software R o número de faltas $x_i = 5$ com frequência $f_i = 0$ como foi apresentado na Tabela 5.4.

Quando existirem muitos valores possíveis para a variável discreta (por exemplo, muitos valores entre 0 e 100) podemos construir uma tabela de distribuição de frequências do mesmo modo que faremos com uma variável contínua (agrupando os dados em classes).

Tabela para dados quantitativos contínuos

Ao contrário das variáveis discretas, as variáveis contínuas assumem, em geral, muitos valores. Isto quer dizer que se utilizássemos para variáveis contínuas tabelas de frequências iguais as tabelas das variáveis discretas teríamos uma tabela com muitas linhas, tornando-a pouco operacional. Para contornar este problema costuma-se descrever as variáveis quantitativas contínuas através de tabelas de frequências cujos valores (x_i) da variável são agrupados em classes (intervalos).

A seguir é apresentado um exemplo de tabela de distribuição de frequências para variáveis contínuas.

■ **Exemplo 5.4** Considere os dados brutos dos salários (em x sal. mín.) de 36 indivíduos:

5,73	13,60	13,23	8,46	17,26	16,22	8,74	23,30	7,39
11,06	13,85	8,12	15,99	10,76	6,26	9,80	5,25	9,77
19,40	10,53	11,59	14,69	8,95	9,35	4,56	4,00	9,13
14,71	12,00	7,59	7,44	6,66	12,79	18,75	6,86	16,61

Na Tabela 5.5 é apresentada a distribuição de frequências dos salários dos 36 indivíduos com os salários agrupados em classes.

Observe que procedendo-se desse modo, ao resumir os dados referentes a uma variável contínua, perde-se alguma informação. Por exemplo, na classe que vai de 12 a 16, não sabemos quais são os valores exatos dos oito salários, a não ser que investiguemos o conjunto de dados original.

Note que estamos usando a notação $a \leftarrow b$ para o intervalo de números contendo o extremo a mas não contendo o extremo b . Podemos também usar a notação $[a, b)$ para designar o mesmo intervalo.

Tabela 5.5: Distribuição de frequências dos salários de 36 indivíduos

Salários	f_i	fr_i	fp_i
4,00 ⊦ 8,00	10	0,2778	27,78%
8,00 ⊦ 12,00	12	0,3333	33,33%
12,00 ⊦ 16,00	8	0,2222	22,22%
16,00 ⊦ 20,00	5	0,1389	13,89%
20,00 ⊦ 24,00	1	0,0278	2,78%
Total	36	1,0000	100,00%

Os comandos do software R para a obtenção da distribuição de frequências do Exemplo 5.4 são apresentados na Lista 5.4.

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3      11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4      19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5      14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Mostrando os dados armazenados (OPCIONAL)
8 dados
9
10 # Organizando os dados em ordem crescente (OPCIONAL)
11 sort(dados)
12
13 # Distribuição de frequências
14 hist(dados, plot=F, breaks=c(4,8,12,16,20,24), right=F)

```

Lista 5.4: Comandos do software R

Observações referentes à Lista 5.4:

- a) no comando `hist()` os limites das classes foram especificados usando o argumento `breaks=c(4,8,12,16,20,24)`;
- b) o padrão do comando `hist()` é gerar intervalos fechados à direita. Para que sejam gerados intervalos abertos à direita deve-se utilizar o argumento `right=F`;
- c) por padrão, o comando `hist()` gera um gráfico (histograma). Como dentro do comando `hist()` foi usado o argumento `plot=F`, ao invés de gerar um gráfico é mostrada uma saída em forma de lista contendo os limites das classe em `$breaks`, as frequências das classes em `$counts`, os pontos médios das classes em `$mids`, etc;

A escolha dos intervalos pode ser feita arbitrariamente e a familiaridade do pesquisador com os dados é que lhe indicará quantas e quais classes devem ser usadas. Quando determinados intervalos de classes tiverem algum significado prático para o pesquisador, ele poderá simplesmente escolher estes intervalos.

Não existe um número "ideal" de classes para um conjunto de dados, embora existam fórmulas para estabelecer quantas classes devem ser construídas. Vejamos, a seguir um critério para a construção das classes.

Passos para construção de uma tabela de distribuição de frequências para dados contínuos

1. Organize os dados em ordem crescente (rol);

2. Determine o número de classes (k):

- (a) Se $n \leq 100$ utilize $k = \sqrt{n}$;
- (b) Se $n > 100$ utilize $k = 1 + 3,22 \log n$;

Observação: Arredonde o valor de k para o valor inteiro mais próximo.

3. Encontre o menor (X_{min}) e o maior (X_{max}) valor do conjunto de dados;

4. Calcule a amplitude (A), que é a diferença entre o maior e o menor valor:

$$A = X_{max} - X_{min}$$

5. Divida a amplitude total pelo número de classes para obter a amplitude das classes (c):

$$c = A/k$$

Observação: Pode ser conveniente arredondar o valor de c para o valor um pouco maior. Por exemplo, se ocorrer $c = 4,21$ arredonde para $c = 4,25$ ou para $c = 4,3$.

6. Organize as classes, iniciando a primeira classe em LI_1 (limite inferior da 1ª classe), em que LI_1 pode ser o menor dado (X_{min}) ou pode ser um valor menor do que ele. Depois basta somar a amplitude das classes (c) a este valor para obter LS_1 (limite superior da 1ª classe), ou seja, $LS_1 = LI_1 + c$. O limite inferior da 2ª classe (LI_2) será igual ao limite superior da 1ª classe, ou seja, $LI_2 = LS_1$. Para obter LS_2 (limite superior da segunda classe) basta somar a amplitude das classes (c) ao limite inferior da mesma classe, ou seja, $LS_2 = LI_2 + c$. Para obter os demais intervalos de classes repita o processo sucessivamente até completar o número de classes (k).

7. Depois de construir todas as (k) classes, conte quantos dados têm dentro de cada classe e anote os resultados na coluna das frequências (f_i).

Observações: (a) A primeira e a última classe não podem ter frequência nula, já nas demais classes pode ocorrer frequência nula. (b) Se ocorrer um dado igual ao limite superior de uma classe este valor deve ser computado na frequência da classe seguinte pois estão sendo considerados intervalos de classes abertos à direita, ou seja, intervalos do tipo $[a, b)$.

■ **Exemplo 5.5** Construiremos uma tabela de distribuição de frequências considerando os dados de salários (x sal. mín.) de 36 indivíduos apresentados no Exemplo 5.4. Para isso, serão seguidos os passos apresentados.

1. Dados em ordem crescente (rol):

4,00	4,56	5,25	5,73	6,26	6,66	6,86	7,39	7,44
7,59	8,12	8,46	8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79	13,23	13,60	13,85
14,69	14,71	15,99	16,22	16,61	17,26	18,75	19,40	23,30

2. Número de classes: $k = \sqrt{n} = \sqrt{36} = 6$;

3. Mínimo e máximo: $X_{min} = 4,00$ e $X_{max} = 23,30$;

4. Amplitude total: $A = X_{max} - X_{min} = 23,30 - 4,00 = 19,30$;

5. Amplitude das classes: $c = A/k = 19,30/6 = 3,22$, que arredondaremos para $c = 3,5$;
6. Classes: iniciaremos o limite inferior da 1ª classe (LI_1) em 4,00 (também poderíamos ter iniciado em um valor menor do que 4,00) e depois iremos somando $c = 3,5$ para obter os limites das demais classes:

$$\begin{aligned} LI_1 &= 4,0 & LS_1 &= 4,0 + c = 4,0 + 3,5 = 7,5 \\ LI_2 &= LS_1 = 7,5 & LS_2 &= 7,5 + c = 7,5 + 3,5 = 11,0 \\ LI_3 &= LS_2 = 11,0 & LS_3 &= 11,0 + c = 11,0 + 3,5 = 14,5 \\ LI_4 &= LS_3 = 14,5 & LS_4 &= 14,5 + c = 14,5 + 3,5 = 18,0 \\ LI_5 &= LS_4 = 18,0 & LS_5 &= 18,0 + c = 18,0 + 3,5 = 21,5 \\ LI_6 &= LS_5 = 21,5 & LS_6 &= 21,5 + c = 21,5 + 3,5 = 25,0 \end{aligned}$$

Na Tabela 5.6 são apresentadas as classes de salários obtidas. Para finalizar a tabela basta observar o conjunto de dados ordenados e contar quantos dados tem em cada classe.

Tabela 5.6: Classes de salários

Salários	f_i
4,0 \leftarrow 7,5	?
7,5 \leftarrow 11,0	?
11,0 \leftarrow 14,5	?
14,5 \leftarrow 18,0	?
18,0 \leftarrow 21,5	?
21,5 \leftarrow 25,0	?
Total	?

7. Frequências: para facilitar a contagem das frequências, neste exemplo, destacamos os dados dentro de cada classe de salários com cores diferentes. **Observação:** não é necessário fazer este procedimento de separar os dados por cores diferentes, foi feito aqui apenas como recurso didático.

4,00	4,56	5,25	5,73	6,26	6,66	6,86	7,39	7,44
7,59	8,12	8,46	8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79	13,23	13,60	13,85
14,69	14,71	15,99	16,22	16,61	17,26	18,75	19,40	23,30

Na Tabela 5.7 pode-se observar a distribuição de frequências dos salários de 36 indivíduos.

■

Observe, neste exemplo, que se ao invés de arredondar a amplitude da classe c para 3,5 tivéssemos arredondado para 4,0 então a sexta classe (última) seria 24 \leftarrow 28 e não teria nenhum elemento; ou seja, na realidade não teríamos seis classes mas apenas cinco (pois a última classe teria frequência igual a zero). Utilizando esta forma de montar as classes precisamos tomar cuidado ao arredondar a amplitude da classe e ao escolher o limite inferior da primeira classe.

Os comandos do software R para a obtenção da distribuição de frequências do Exemplo 5.5 são apresentados na Lista 5.5.

Tabela 5.7: Distribuição de frequências dos salários de 36 indivíduos

Salários	f_i
4,0 \vdash 7,5	9
7,5 \vdash 11,0	11
11,0 \vdash 14,5	7
14,5 \vdash 18,0	6
18,0 \vdash 21,5	2
21,5 \vdash 25,0	1
Total	36

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3     11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4     19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5     14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Distribuição de frequências
8 hist(dados, plot=F, breaks=c(4,7.5,11,14.5,18,21.5,25), right=F)

```

Lista 5.5: Comandos do software R

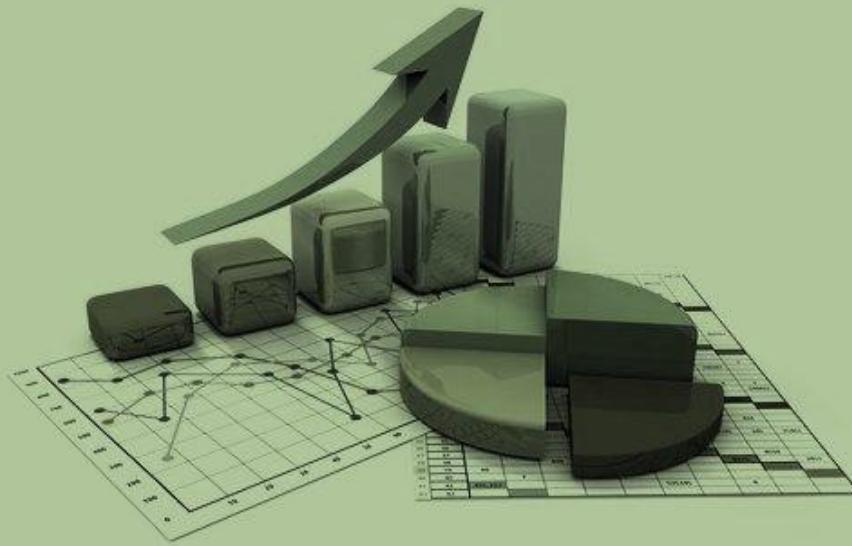
Pode-se, ainda, agrupar os dados em uma distribuição de frequências deixando que o software R utilize um critério próprio para escolha do número de classes e das próprias classes, sendo que para isso basta não utilizar o argumento `breaks=c(...)` dentro do comando `hist()`. Na Lista 5.6 são apresentados os comandos do R para obtenção de uma distribuição de frequências usando o critério do R para obtenção das classes.

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3     11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4     19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5     14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Distribuição de frequências
8 hist(dados, plot=F, right=F)

```

Lista 5.6: Comandos do software R



6. Gráficos

6.1 Gráficos de colunas e barras

6.1.1 Gráfico de colunas (ou gráfico de barras verticais)

O gráfico de colunas é um tipo de gráfico de barras, porém, nesse caso as barras são verticais (colunas).

Passos para construir um gráfico de colunas

1. desenhe o sistema de eixos cartesianos;
2. anote as categorias da variável estudada no eixo das abscissas (eixo horizontal);
3. escreva as frequências ou as frequências relativas (proporções) ou as frequências percentuais (porcentagens) no eixo das ordenadas (eixo vertical), obedecendo uma escala;
4. desenhe barras verticais (colunas) de mesma largura, separadas por um espaço, para representar as categorias da variável em estudo. A altura de cada barra deve ser dada pela frequência da categoria;
5. coloque legenda nos dois eixos (nomes dos eixos) e título na figura.

■ **Exemplo 6.1** No Exemplo 5.2 foi construída uma tabela de distribuição de frequências para a variável grau de instrução (Tabela 5.2). Esta tabela é apresentada novamente para facilitar a construção do gráfico de colunas.

Recordando a Tabela 5.2

Grau de instrução	f_i
Fundamental	12
Médio	18
Superior	6
Total	36

Podemos representar graficamente esta tabela de distribuição de frequências por meio de um gráfico de colunas (Figura 6.1).

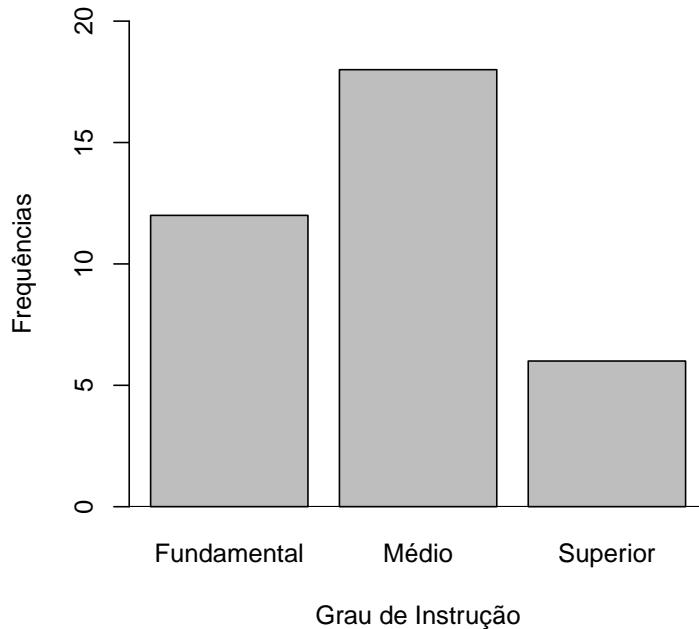


Figura 6.1: Gráfico de colunas do grau de instrução

Na Lista 6.1 são apresentados os comandos do software R para fazer o gráfico de colunas (ou barras verticais).

```

1 # Entrando com os dados no R:
2 dados <- c("Fundamental", "Médio", "Médio", "Superior", "Médio", "Médio",
3         "Médio", "Médio", "Fundamental", "Fundamental", "Superior", "Fundamental",
4         "Superior", "Fundamental", "Médio", "Médio", "Médio", "Médio",
5         "Fundamental", "Médio", "Superior", "Fundamental", "Superior", "Médio",
6         "Médio", "Fundamental", "Médio", "Médio", "Fundamental",
7         "Fundamental", "Fundamental", "Superior", "Fundamental", "Médio")
8
9 # Mostrando os dados armazenados (OPCIONAL)
10 dados
11
12 # Gerando a distribuição de frequências
13 tab <- table(dados)
14
15 # Mostrando a distribuição de frequências (OPCIONAL)
16 tab
17
18 # Gerando o gráfico de colunas
19 barplot(tab, ylim=c(0,20), xlab="Grau de Instrução", ylab="Frequências")
20 abline(h=0)
21
22 # Gerando o gráfico de colunas com título
23 barplot(tab, ylim=c(0,20), xlab="Grau de Instrução", ylab="Frequências",
24         main="Gráfico de colunas do grau de instrução")
25 abline(h=0)
```

Lista 6.1: Comandos do software R

Observações referentes à Lista 6.1:

- a) o argumento `ylim` especifica os limites do eixo `y` (eixo das ordenadas). Assim, `ylim=c(0, 20)` especifica que o eixo `y` inicia em $y = 0$ e termina em $y = 20$;
- b) os argumentos `xlab` e `ylab` especificam os nomes dos eixos `x` e `y`, respectivamente;
- c) o comando `abline (h=0)` faz uma linha horizontal passando por $y = 0$.

6.1.2 Gráfico de barras (ou gráfico de barras horizontais)

O gráfico de barras (ou gráfico de barras horizontais) é usado para apresentar variáveis qualitativas, sejam elas nominais ou ordinais. Um gráfico de barras é a representação gráfica de uma distribuição de frequências de dados qualitativos.

Passos para construir um gráfico de barras

1. desenhe o sistema de eixos cartesianos;
2. anote as categorias da variável estudada no eixo das ordenadas (eixo vertical);
3. escreva as frequências ou as frequências relativas (proporções) ou as frequências percentuais (porcentagens) no eixo das abscissas (eixo horizontal), obedecendo uma escala;
4. desenhe barras horizontais de mesma largura, separadas por um espaço, para representar as categorias da variável em estudo. O comprimento de cada barra deve ser dado pela frequência da categoria;
5. coloque legenda nos dois eixos (nomes dos eixos) e título na figura.

■ **Exemplo 6.2** No Exemplo 5.2 foi construída uma tabela de distribuição de frequências para a variável grau de instrução (Tabela 5.2). Podemos representar graficamente esta tabela de distribuição de frequências por meio de um gráfico de barras (Figura 6.2).

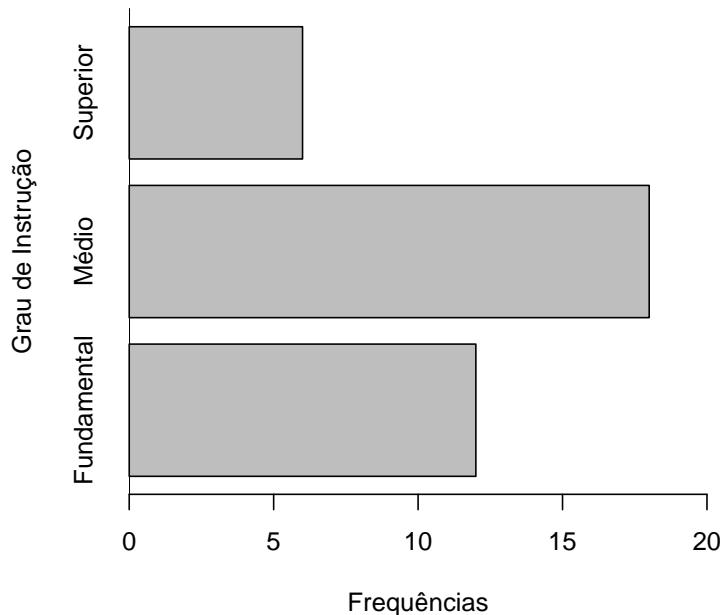


Figura 6.2: Gráfico de barras do grau de instrução

■ Na Lista 6.2 são apresentados os comandos do software R para fazer o gráfico de barras horizontais.

```

1 # Entrando com os dados no R:
2 dados <- c("Fundamental", "Médio", "Médio", "Superior", "Médio", "Médio",
3         "Médio", "Médio", "Fundamental", "Fundamental", "Superior", "Fundamental",
4         "Superior", "Fundamental", "Médio", "Médio", "Médio",
5         "Fundamental", "Médio", "Superior", "Fundamental", "Superior", "Médio",
6         "Médio", "Fundamental", "Médio", "Médio", "Fundamental",
7         "Fundamental", "Fundamental", "Superior", "Fundamental", "Médio", "Médio")
8
9 # Mostrando os dados armazenados (OPCIONAL)
10 dados
11
12 # Gerando a distribuição de frequências
13 tab <- table(dados)
14
15 # Mostrando a distribuição de frequências (OPCIONAL)
16 tab
17
18 # Gerando o gráfico de barras horizontais
19 barplot(tab, horiz=T, xlim=c(0,20), xlab="Frequências",
20         ylab="Grau de Instrução")
21 abline(v=0)

```

Lista 6.2: Comandos do software R

Observações referentes à Lista 6.2:

- a) o argumento `horiz=T` faz com que as barras sejam geradas na posição horizontal;
- b) o argumento `xlim` especifica os limites do eixo x (eixo das abscissas). Assim, `xlim=c(0, 20)` especifica que o eixo x inicia em $x = 0$ e termina em $x = 20$;
- c) o argumento `abline (v=0)` faz uma linha vertical passando por $x = 0$.

6.2 Gráfico de setores

O gráfico de setores (ou gráfico de pizza) é especialmente indicado para apresentar variáveis qualitativas, desde que o número de categorias seja pequeno.

Passos para construir um gráfico de setores

1. desenhe uma circunferência (com 360°). Essa circunferência representará o total, ou seja, 100% dos dados;
2. divida a circunferência em tantos setores quantas sejam as categorias da variável em estudo, mas é preciso calcular o ângulo de cada setor. O ângulo é igual a proporção da categoria multiplicada por 360° ;
3. marque na circunferência os ângulos calculados e desenhe os raios passando pelos ângulos;
4. pinte cada categoria de uma cor, escreva as legendas das categorias e coloque título na figura.

■ **Exemplo 6.3** No Exemplo 5.2 foi construída uma tabela de distribuição de frequências para a variável grau de instrução (Tabela 5.2). Podemos representar graficamente esta tabela de distribuição de frequências por meio de um gráfico de setores(Figura 6.3).

Para construir o gráfico de setores calculamos as proporções de cada categoria e depois multiplicamos por 360° para encontrar o ângulo de cada setor:

Fundamental	$12/36 = 0,3333$	$0,3333 \times 360^\circ \approx 120^\circ$
Médio	$18/36 = 0,5000$	$0,5000 \times 360^\circ = 180^\circ$
Superior	$6/36 = 0,1667$	$0,1667 \times 360^\circ \approx 60^\circ$

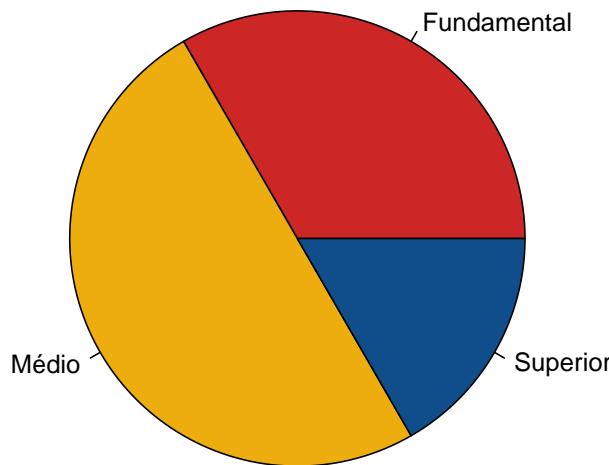


Figura 6.3: Gráfico de setores do grau de instrução

Observe que o setor do ensino fundamental tem um ângulo de 120° , o setor do ensino médio tem um ângulo de 180° e o setor do ensino superior tem um ângulo de 60° .

Outra forma de apresentação mais elegante do gráfico de setores é utilizando um gráfico 3D. Na Figura 6.4 é apresentado um gráfico de setores 3D feito com o pacote `plotrix` do software R.

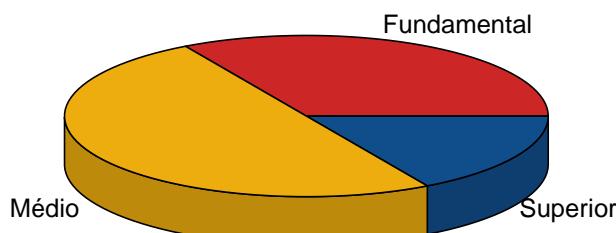


Figura 6.4: Gráfico de setores 3D do grau de instrução

■

Na Lista 6.3 são apresentados os comandos do software R para fazer os gráficos de setores 2D (Figura 6.3) e 3D (Figura 6.4).

```

1 # Entrando com os dados no R:
2 dados <- c("Fundamental", "Médio", "Médio", "Superior", "Médio", "Médio",
3      "Médio", "Médio", "Fundamental", "Fundamental", "Superior", "Fundamental",
4      "Superior", "Fundamental", "Médio", "Médio", "Médio", "Médio",
5      "Fundamental", "Médio", "Superior", "Fundamental", "Superior", "Médio",

```

```

6      "Médio", "Fundamental", "Médio", "Médio", "Médio", "Fundamental",
7      "Fundamental", "Fundamental", "Superior", "Fundamental", "Médio", "Médio")
8
9  # Mostrando os dados armazenados (OPCIONAL)
10 dados
11
12 # Gerando a distribuição de frequências
13 tab <- table(dados)
14
15 # Mostrando a distribuição de frequências (OPCIONAL)
16 tab
17
18 # Gráfico de setores 2D
19 pie(tab, col=c("firebrick3", "darkgoldenrod2", "dodgerblue4"))
20
21
22 #Gráfico de setores 3D
23
24 # Carregando o pacote plotrix (precisa estar instalado)
25 library(plotrix)
26
27 # Plotando o gráfico:
28 pie3D(tab, col=c("firebrick3", "darkgoldenrod2", "dodgerblue4"),
29       labels=names(tab), labelcex=1)

```

Lista 6.3: Comandos do software R

Observe, na Lista 6.3, que dentro dos comandos usados para gerar os gráficos de setores (`pie()` e `pie3D()`) foi utilizado o argumento `col`, que serve para especificar as cores do gráfico. Neste exemplo foram usadas as cores `"firebrick3"`, `"darkgoldenrod2"` e `"dodgerblue4"`. Se digitar o comando `colors()` no R irá aparecer uma lista com os nomes de mais de 600 cores (porém sem mostrar as cores). Para observar diferentes cores e seus respectivos nomes para usá-las no software R pode-se consultar uma paleta de cores. Em uma consulta rápida na internet podem ser encontradas várias paletas de cores para o software R. Uma paleta de cores interessante pode ser acessada em: <http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>. Outra forma de visualizar exemplos de cores no R é digitando o comando `demon(colors)` diretamente no console do R.

Além da possibilidade de alterar as cores também é possível fazer alterações em outros parâmetros gráficos. Para visualizar os parâmetros gráficos que podem ser alterados basta digitar no console `?pie` ou `?pie3D`, ou seja, o nome do comando antecedido pelo sinal de interrogação.

6.3 Diagrama de linhas

O diagrama de linhas é utilizado para apresentar graficamente dados quantitativos discretos organizados em uma tabela de distribuição de frequências.

Passos para construir um diagrama de linhas

1. desenhe o sistema de eixos cartesianos;
2. no eixo das abscissas (eixo horizontal) apresente uma escala (de um em um, ou de dois em dois, ...) de valores para a variável;
3. no eixo das ordenadas (eixo vertical) apresente uma escala de valores para as frequências;
4. desenhe linhas verticais a partir dos valores assumidos pela variável no eixo das abscissas.
Os comprimentos das barras são dados pelas frequências ou pelas porcentagens;
5. coloque legendas nos dois eixos e título na figura.

■ **Exemplo 6.4** No Exemplo 5.3 foi construída uma tabela de distribuição de frequências para a variável número de faltas ao trabalho (Tabela 5.4). Esta tabela é apresentada novamente para facilitar a construção do diagrama de linhas.

Recordando a Tabela 5.4

Número de faltas	f_i
0	9
1	10
2	5
3	3
4	2
5	0
6	1
Total	30

Podemos representar essa distribuição de frequências por meio de um diagrama de linhas (Figura 6.5).

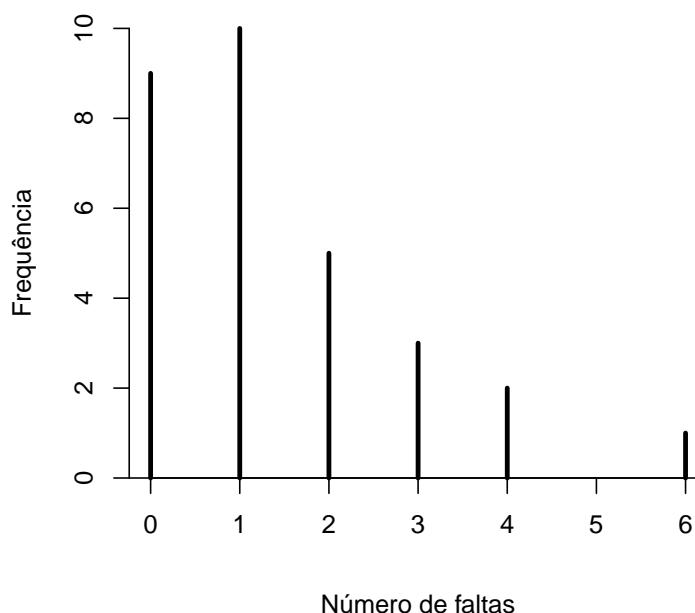


Figura 6.5: Diagrama de linhas do número de faltas

■

Na Lista 6.4 são apresentados os comandos do software R para fazer o diagrama de linhas (Figura 6.5).

```

1 # Entrando com os dados no R
2 dados <- c(1, 3, 1, 1, 0, 1, 0, 1, 1, 0, 2, 2, 0, 0, 0, 1, 2, 1, 2, 0, 0, 1,
   6, 4, 3, 3, 1, 2, 4, 0)

```

```

3 # Agrupando os dados em uma distribuição de frequências
4 tab <- table(dados)
5
6
7 # Plotando o diagrama de linhas
8 plot(tab, xlab="Número de faltas", ylab="Frequência", lwd=3, axes=F,
9     frame.plot=F)
10 axis(1, pos=0); axis(2); abline(h=0)

```

Lista 6.4: Comandos do software R

Observações referentes à Lista 6.4:

- dentro do comando `plot()` foram alterados alguns parâmetros gráficos. Foi usado o argumento `lwd=3` para aumentar a expressura das linhas; foi usado o argumento `axes=F` para omitir os eixos do gráfico; e foi usado o argumento `frame.plot=F` para omitir a "caixa" ao redor do gráfico;
- o comando `axis(1, pos=0)` exibe o eixo das abcissas (eixo 1), e o argumento `pos=0` especifica que este eixo passará pela posição $y = 0$;
- o comando `axis(2)` exibe o eixo das ordenadas (eixo 2).

6.4 Histograma

O histograma é utilizado para apresentar graficamente dados quantitativos contínuos agrupados em uma tabela de distribuição de frequências.

Passos para construir um histograma

- desenhe o sistema de eixos cartesianos;
- no eixo das abscissas (eixo horizontal) apresente os limites das classes;
- no eixo das ordenadas (eixo vertical) apresente uma escala de valores para as frequências;
- desenhe barras com alturas iguais às frequências (ou às proporções ou porcentagens) das respectivas classes. As barras devem ser justapostas (não são separadas), a fim de evidenciar a natureza contínua da variável;
- coloque legendas nos dois eixos e título na figura.

■ **Exemplo 6.5** No Exemplo 5.4 foi construída uma tabela de distribuição de frequências para a variável salário (Tabela 5.5). Esta tabela é apresentada novamente para facilitar a construção do histograma.

Recordando a Tabela 5.5

Salários	f_i	fr_i	fp_i
4,00 ⊜ 8,00	10	0,2778	27,78%
8,00 ⊜ 12,00	12	0,3333	33,33%
12,00 ⊜ 16,00	8	0,2222	22,22%
16,00 ⊜ 20,00	5	0,1389	13,89%
20,00 ⊜ 24,00	1	0,0278	2,78%
Total	36	1,0000	100,00%

Podemos representar essa distribuição de frequências por meio de um histograma (Figura 6.6).

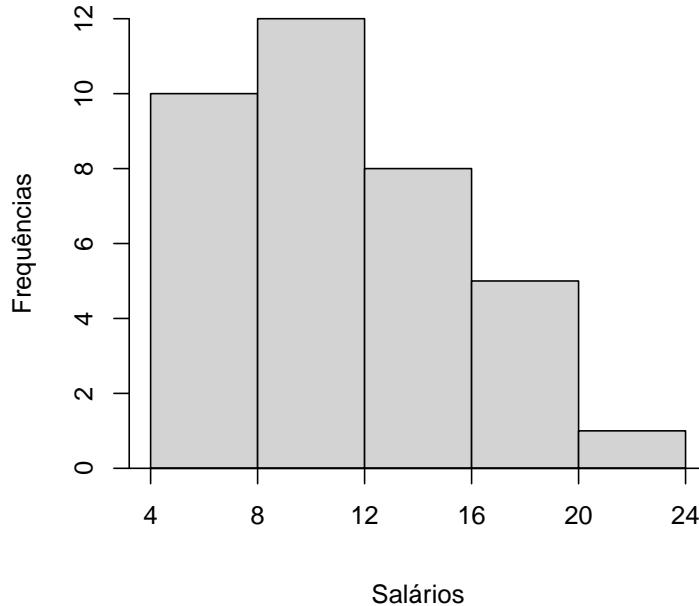


Figura 6.6: Histograma dos salários

■

Na Lista 6.5 são apresentados os comandos do software R para fazer o histograma (Figura 6.6).

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3      11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4      19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5      14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Plotando o histograma
8 hist(dados, breaks=c(4,8,12,16,20,24), xlab="Salários", ylab="Frequências",
9      right=F, axes=F, col="lightgray", main="")
10 axis(1,c(4,8,12,16,20,24), pos=0); axis(2); abline(h=0)
11
12 # Plotando um outro histograma sem especificar as classes
13 hist(dados, xlab="Salários", ylab="Frequências", right=F, col="lightgray",
14      main="")

```

Lista 6.5: Comandos do software R

Observações referentes à Lista 6.5:

- no comando `hist()` os limites das classes foram especificados usando o argumento `breaks=c(4,8,12,16,20,24)`;
- o padrão do comando `hist()` é gerar intervalos fechados à direita. Para que sejam gerados intervalos abertos à direita deve-se utilizar o argumento `right=F`;
- o argumento `axes=F` faz com que os eixos (x e y) não sejam gerados;
- no argumento `col` pode ser especificada a cor do gráfico;
- quando o argumento `breaks` não é informado dentro do comando `hist()` o software R gera classes usando um critério próprio, diferente do que foi visto neste material.

6.5 Polígono de frequências

Dados quantitativos contínuos organizados em uma tabela de distribuição de frequências também podem ser apresentados em polígonos de frequências.

O polígono de frequências utiliza os pontos médios das classes. Para calcular o ponto médio (x_i) da classe i basta calcular a média entre o limite inferior e o limite superior desta classe, ou seja,

$$x_i = \frac{LI_i + LS_i}{2}.$$

Passos para construir um polígono de frequências

1. desenhe o sistema de eixos cartesianos;
2. apresente os pontos médios das classes no eixo das abscissas (eixo horizontal);
3. apresente as frequências no eixo das ordenadas (eixo vertical);
4. marque pontos de coordenadas (x_i, f_i) em que a coordenada x_i é o ponto médio da classe e a coordenada f_i é a frequência da classe;
5. una os pontos por segmentos de reta;
6. feche o polígono unindo os extremos da figura com o eixo horizontal. Utilize duas classes auxiliares, uma antes da primeira classe e outra depois da última classe, ambas com frequência zero, e utilize os pontos médios dessas classes auxiliares para fechar o polígono;
7. coloque legendas nos dois eixos e título na figura.

■ **Exemplo 6.6** No Exemplo 5.4 foi construída uma tabela de distribuição de frequências para a variável salário (Tabela 5.5). Podemos representar essa distribuição de frequências por meio de um polígono de frequências (Figura 6.7).

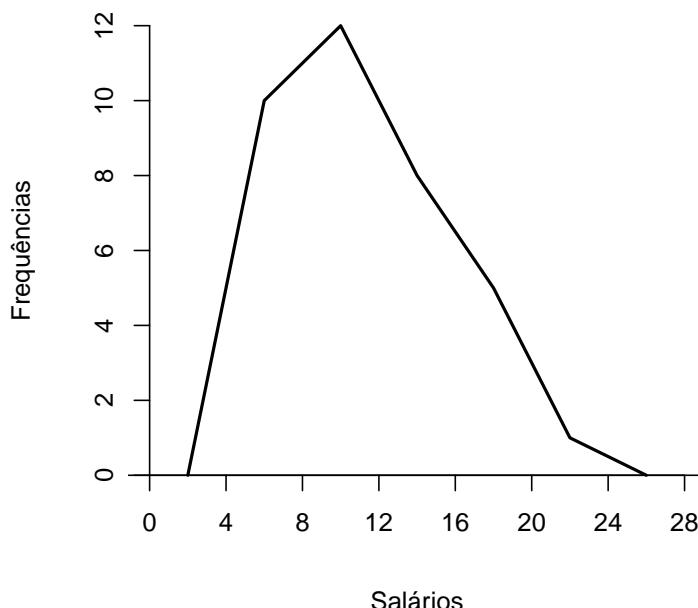


Figura 6.7: Polígono de frequências dos salários

Para construir o polígono de frequências deve-se "criar" duas classes auxiliares, uma antes da primeira classe e outra depois da última classe, ambas com frequência zero, e obter os pontos médios dessas classes, como pode ser observado na Tabela 6.4.

Tabela 6.4: Tabela auxiliar para construção do polígono de frequências

Salários	Ponto médio (x_i)	Frequência (f_i)
0,00 ⊜ 4,00	2	0
4,00 ⊜ 8,00	6	10
8,00 ⊜ 12,00	10	12
12,00 ⊜ 16,00	14	8
16,00 ⊜ 20,00	18	5
20,00 ⊜ 24,00	22	1
24,00 ⊜ 28,00	26	0

Depois de construir a tabela auxiliar marque os pontos (x_i, f_i) (Figura 6.8A) e depois ligue os pontos usando segmentos de retas (Figura 6.8B).

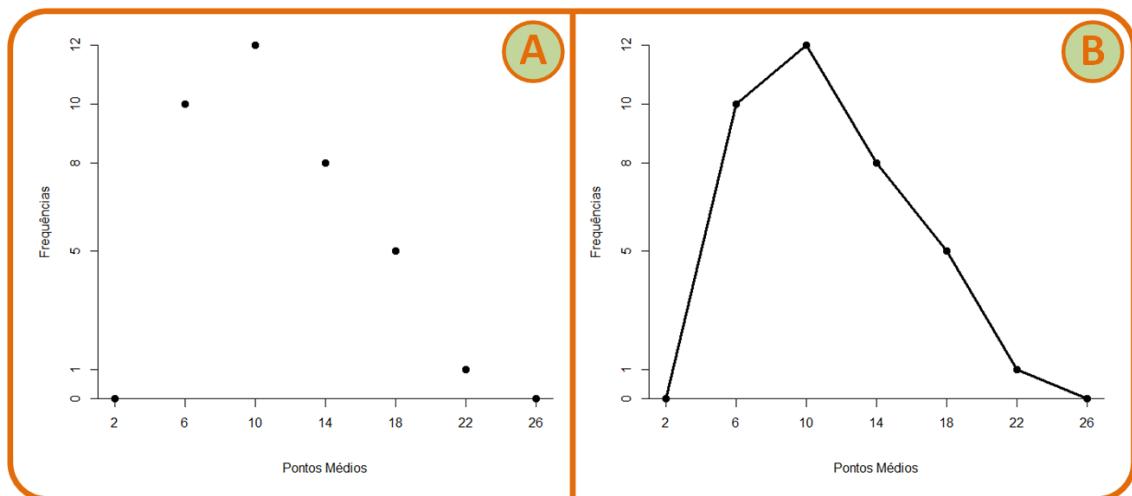


Figura 6.8: Construção do polígono de frequências

Na Lista 6.6 são apresentados os comandos do software R para fazer o polígono de frequências (Figura 6.7).

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3      11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4      19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5      14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Agrupando os dados em classes
8 hg <- hist(dados, breaks=c(4,8,12,16,20,24), plot=F, right=F)
9
10 # Mostrando o objeto hg
11 hg
12
13 # Amplitude das classes
14 c <- diff(hg$breaks[1:2])

```

```

15 # Pontos médios das classes (com classes auxiliares)
16 pm = c(hg$mids[1]-c, hg$mids, hg$mids[length(hg$mids)]+c)
17
18 # Frequências das classes (classes auxiliares freq. zero)
19 freq <- c(0, hg$counts, 0)
20
21 # Plotando o polígono de frequências
22 plot(pm, freq, type="l", lwd=2, bty="l", xlab="Salários",
23      ylab="Frequências", main="", axes=F, xlim=c(0,max(hg$breaks)+c))
24 axis(1, seq(0,max(hg$breaks)+c,c), pos=0); axis(2, pos=0); abline(h=0)
25

```

Lista 6.6: Comandos do software R

Observações referentes à Lista 6.6:

- primeiramente foi criado um objeto (`hg`) contendo a distribuição de frequências. Esse objeto é do tipo lista e na saída desse objeto são mostradas várias estruturas de dados. Por exemplo, em `hg$breaks` são dados os limites das classes, em `hg$counts` são dadas as frequências das classes e em `hg$mids` são dados os pontos médios das classes, etc;
- a amplitude das classes (`c`) foi obtida pela diferença (comando `diff()`) entre os dois primeiros limites das classes, ou seja, `diff(hg$breaks[1:2])`;
- os pontos médios das classes são obtidos no objeto `hg` em `$mids`. Para obter os pontos médios das classes auxiliares a amplitude das classes (`c`) foi subtraída da primeira classe (`hg$mids[1]-c`) e somada à última classe (`hg$mids[length(hg$mids)]+c`);
- foi atribuído valor zero às frequências das classes auxiliares e foram mantidas as frequências das classes originais usando o comando `c(0, hg$counts, 0)` para criar o objeto `freq`.

Também é possível apresentar o histograma e o polígono de frequências sobrepostos em um mesmo gráfico. Na Lista 6.7 são apresentados os comandos do software R para fazer o histograma e o polígono de frequências juntos (Figura 6.9).

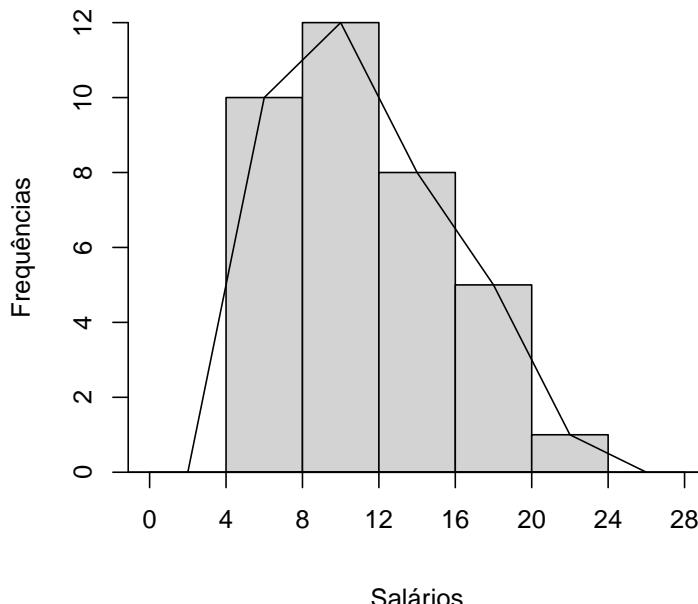


Figura 6.9: Histograma com polígono de frequências

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3      11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4      19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5      14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Pacote agricolae
8 library(agricolae)
9
10 # Histograma (deve-se criar um objeto)
11 histograma <- hist(dados, breaks=c(4,8,12,16,20,24), xlim=c(0, 28),
12      xlab="Salários", ylab="Frequências", right=F, axes=F,
13      col="lightgray", main="")
14 axis(1,c(0,4,8,12,16,20,24,28), pos=0); axis(2); abline(h=0)
15
16 # Polígono de frequências sobreposto ao histograma
17 polygon.freq(histograma)

```

Lista 6.7: Comandos do software R

6.6 Ramo e folhas

Um "gráfico" simples para resumir um conjunto de dados quantitativos com poucas observações é o ramo-e-folhas. Uma vantagem do ramo-e-folhas é que ele explicita os dados. Não existe uma regra fixa para construir o ramo-e-folhas, mas a idéia básica é dividir cada observação em duas partes: a primeira, o ramo, é colocada à esquerda de uma linha vertical, a segunda, as folhas, é colocada à direita.

O ramo e as folhas podem representar quaisquer grandezas: unidades e decimais, dezenas e unidades, centenas e dezenas, etc. Por esse motivo é aconselhável que o ramo-e-folhas tenha uma legenda para facilitar sua interpretação.

■ **Exemplo 6.7** Considere as idades, em meses, de 30 animais:

82	72	22	44	30	41	32	78	43	22	65	56	42	65	61
61	71	88	52	55	101	24	62	50	68	80	55	12	18	32

Os dados podem ser organizados, separando-os pela dezenas, uma em cada linha:

12	18				
22	22	24			
30	32	32			
41	42	43	44		
50	52	55	55	56	
61	61	62	65	65	68
71	72	78			
80	82	88			
101					

Os dados em cada linha tem as dezenas em comum, logo pode-se colocar as dezenas em evidência, separando-as das unidades por uma linha vertical. Ao dispor os dados dessa forma, é construído um diagrama de ramo-e-folhas (Figura 6.10). ■

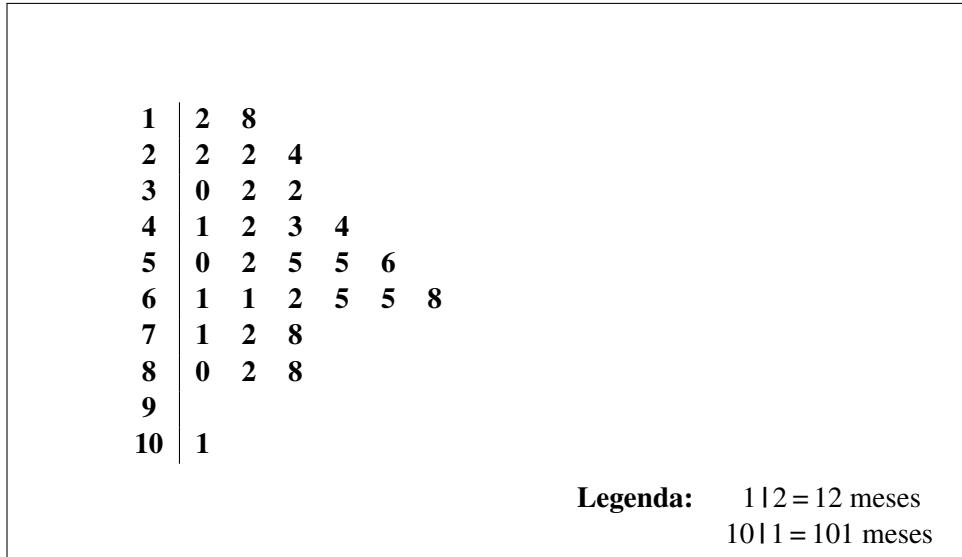


Figura 6.10: Ramo-e-folhas das idades

Na Lista 6.8 são apresentados os comandos do software R para fazer o ramo-e-folhas das idades.

```

1 # Entrando com os dados no R
2 dados <- c(82, 72, 22, 44, 30, 41, 32, 78, 43, 22, 65, 56, 42, 65, 61,
3   61, 71, 88, 52, 55, 101, 24, 62, 50, 68, 80, 55, 12, 18, 32)
4
5 # Ramo-e-folhas
6 stem(dados)

```

Lista 6.8: Comandos do software R

- **Exemplo 6.8** Considere os dados brutos dos salários (em x sal. mín.) de 36 indivíduos:

5,73	13,60	13,23	8,46	17,26	16,22	8,74	23,30	7,39
11,06	13,85	8,12	15,99	10,76	6,26	9,80	5,25	9,77
19,40	10,53	11,59	14,69	8,95	9,35	4,56	4,00	9,13
14,71	12,00	7,59	7,44	6,66	12,79	18,75	6,86	16,61

Esses dados podem ser representados em um ramo-e-folhas em que a unidade representa os ramos e a parte decimal representa as folhas (Figura 6.11). **Observação:** deve-se arredondar os dados para uma casa decimal para que cada valor à direita da linha vertical represente uma folha (cada número seja referente a um dado).

Na Lista 6.9 são apresentados os comandos do software R para fazer o ramo-e-folhas para os dados de salários.

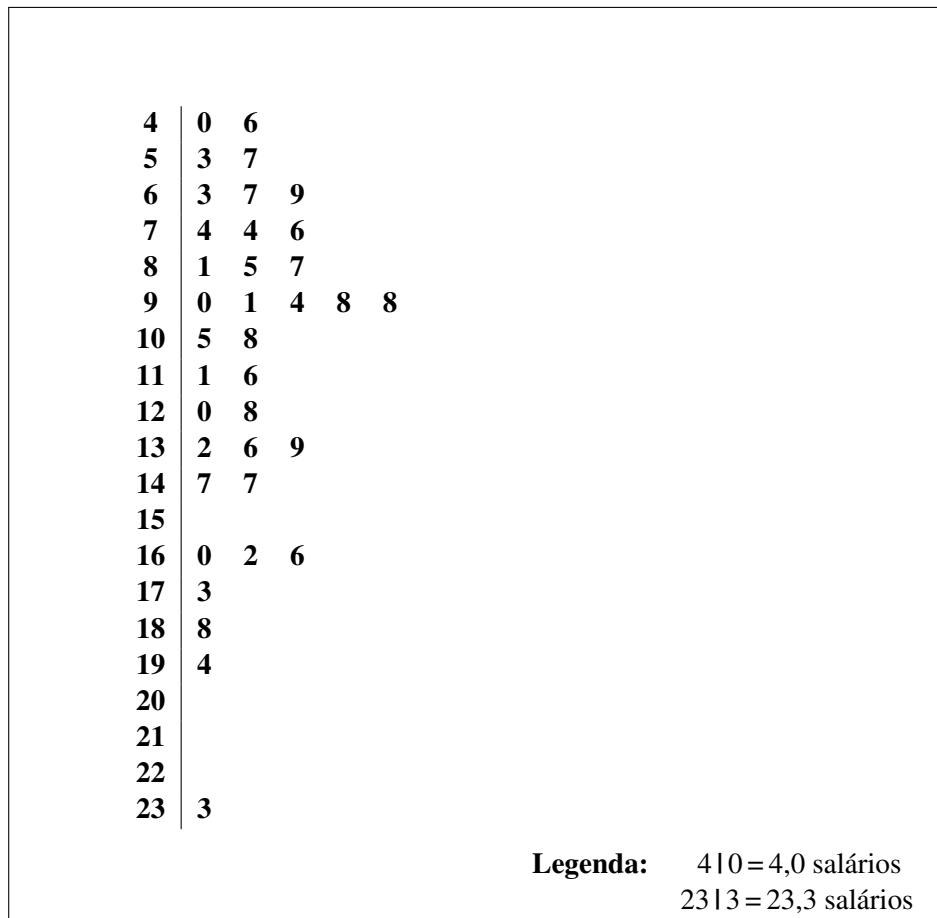


Figura 6.11: Ramo-e-folhas dos salários

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3      11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4      19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5      14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Arredondando os dados para uma casa decimal e ordenando(OPCIONAL)
8 sort(round(dados,1))
9
10 # Ramo-e-folhas
11 stem(dados, scale=2)

```

Lista 6.9: Comandos do software R

Observação: Na Lista 6.9, no comando `stem()`, o parâmetro `scale` foi alterado para `scale=2` para que fossem exibidos todos os valores do ramo (de 4 à 23).

6.7 Gráfico de Série Temporal

Este gráfico serve para representar uma série temporal, ou seja, dados coletados em diferentes momentos do tempo.

Passos para construir um gráfico de série temporal

1. desenhe o sistema de eixos cartesianos;
2. anote os períodos de tempo da variável estudada no eixo das abscissas (eixo horizontal);
3. escreva as frequências ou taxas no eixo das ordenadas (eixo vertical), obedecendo uma escala;
4. marque pontos de coordenadas $(x_i; f_i)$ em que a coordenada x_i representa o tempo no qual foi registrada a informação e a coordenada f_i é a frequência ou taxa observada no respectivo tempo;
5. una os pontos por segmentos de reta;
6. coloque legenda nos dois eixos (nomes dos eixos) e título na figura;
7. no caso de existir mais de uma série, coloque uma legenda no gráfico para identificar cada uma das séries.

■ **Exemplo 6.9** A Tabela 6.5 apresenta informações sobre a taxa de ocupação da rede hoteleira dos municípios de Búzios e Petrópolis, no estado do Rio de Janeiro, no período de 2000 a 2008.

Tabela 6.5: Taxa de ocupação da rede hoteleira

Ano	Búzios	Petrópolis
2000	72.8	60.6
2001	66.2	53.7
2002	69.2	55.3
2003	65.9	56.7
2004	62.4	56.4
2005	67.8	57.8
2006	61.3	57.5
2007	68.5	59.8
2008	70.4	63.3

Podemos representar graficamente esta tabela como um gráfico de duas séries temporais (Figura 6.12). ■

Na Lista 6.10 são apresentados os comandos do software R para fazer o gráfico de séries temporais (Figura 6.12).

```

1 # Entrando com os dados
2 Ano <- 2000:2008
3 Buzios <- c(72.8, 66.2, 69.2, 65.9, 62.4, 67.8, 61.3, 68.5, 70.4)
4 Petropolis <- c(60.6, 53.7, 55.3, 56.7, 56.4, 57.8, 57.5, 59.8, 63.3)
5
6 # Gerando o gráfico de séries temporais de Búzios
7 plot(Ano, Buzios, type="l", xlab="Ano", ylab="Taxa de ocupação %",
8      xlim=c(2000,2008), ylim=c(50,80), col="blue", lwd=2, pch=16)
9
10 # Sobrepondo a série de Petropólis ao gráfico de Búzios gerado anteriormente
11 lines(Ano, Petropolis, col="red", lwd=2)
12
13 # Adicionando a legenda
14 legend("topright", c("Búzios", "Petrópolis"), lwd=2, col=c("blue", "red"))

```

Lista 6.10: Comandos do software R

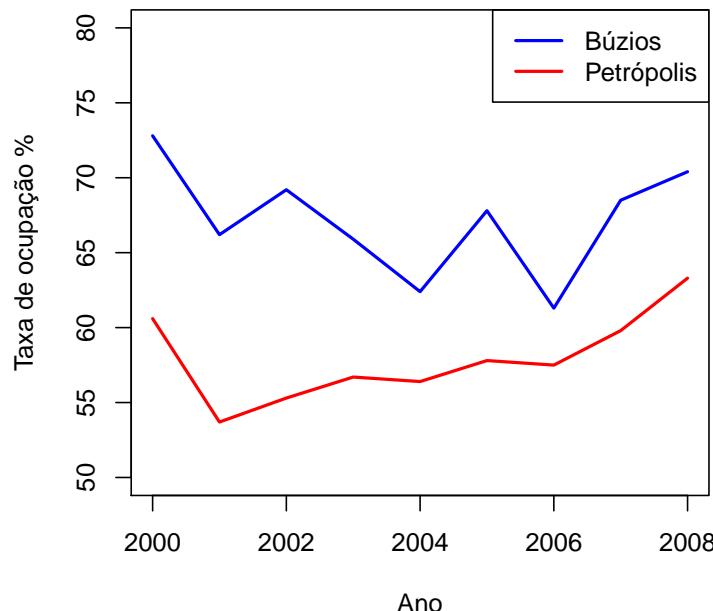
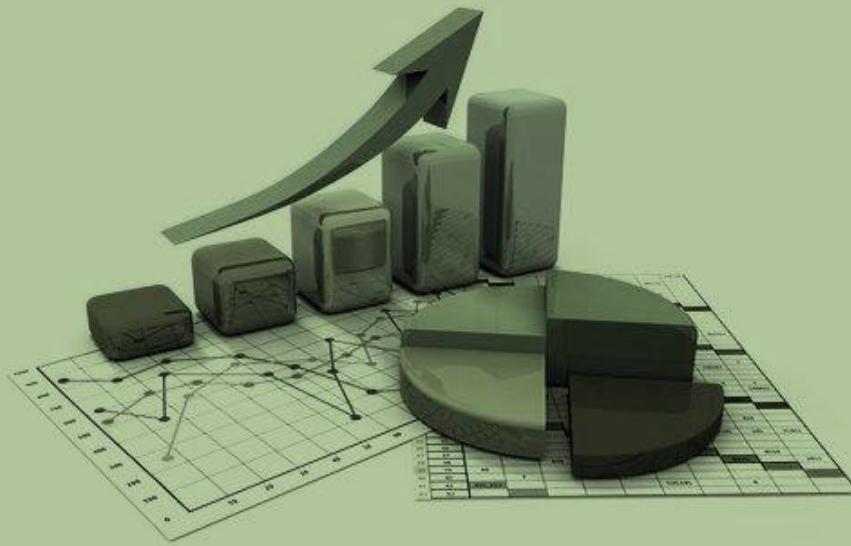


Figura 6.12: Taxa de ocupação hoteleira

Observações referentes à Lista 6.10:

- a) Foi criado o objeto `ano` para representar o período de tempo no qual foram coletadas as informações;
- b) Foram criados os objetos `Buzios` e `Petropolis` contendo as taxas de ocupação da rede hoteleira observadas em cada ano de acordo com cada município;
- c) No comando `plot()` é especificado o objeto relativo à primeira série estudada (`Búzios`);
- d) O argumento `type="l"` indica que o gráfico será realizado no formato de uma linha;
- e) O argumento `lwd=2` indica a espessura da linha do gráfico;
- f) No comando `lines()` é especificado o objeto relativo à segunda série estudada (`Petrópolis`);
- g) No comando `legend()` são especificados: o local do gráfico (coordenadas) no qual a legenda será plotada, o vetor com os nomes de cada série, a espessura e as cores das linhas.



7. Medidas de posição

7.1 Introdução

Foi visto nas seções anteriores que o resumo de dados por meio de tabelas de distribuição de frequências e gráficos fornece muito mais informação sobre o comportamento de uma variável do que o próprio conjunto original de dados (dados brutos). Muitas vezes, é necessário resumir ainda mais os dados, apresentando algumas medidas que sejam capazes de resumir aspectos importantes da distribuição da variável de interesse. Usualmente, emprega-se uma das seguintes *medidas de posição*: média, mediana ou moda. Essas medidas de posição também são denominadas de *medidas de tendência central*.

7.2 Somatório

7.2.1 Variáveis e índices

O símbolo x_i (leia x índice i) representa qualquer um dos n valores x_1, x_2, \dots, x_n assumidos por uma variável aleatória X no conjunto de dados. A letra i , usada como índice, indica a "posição" (de 1 a n) do elemento x no conjunto de dados. Assim, x_1 é o elemento que ocupa a primeira posição na amostra, x_2 é o elemento que ocupa a segunda posição na amostra e assim consecutivamente até x_n que é o elemento que ocupa a n -ésima posição na amostra.

■ **Exemplo 7.1** Se for considerada uma amostra de tamanho $n = 3$ pessoas e se X representa uma variável relativa ao peso, em kg , então uma possibilidade de resultados é:

$$50,5, 64,3 \text{ e } 72,6$$

Logo, $x_1 = 50,5$, $x_2 = 64,3$ e $x_3 = 72,6$. ■

7.2.2 Notação de somatório

Para representarmos a soma de n variáveis aleatórias podemos utilizar o símbolo Σ , letra grega maiúscula sigma (Equação 7.1).

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n \quad (7.1)$$

ou seja, a soma $x_1 + x_2 + \dots + x_n$ pode ser representada por $\sum_{i=1}^n x_i$. Lê-se somatório de x_i para i variando de 1 até n .

■ **Exemplo 7.2** Considere os dados de peso de $n = 3$ pessoas: $\{50,5, 64,3 \text{ e } 72,6\}$.

O somatório desses dados é calculado por:

$$\begin{aligned} \sum_{i=1}^3 x_i &= x_1 + x_2 + x_3 \\ &= 50,5 + 64,3 + 72,6 \\ &= 187,4 \end{aligned}$$

■

Na Lista 7.1 são apresentados os comandos do software R para realizar a entrada dos dados e o cálculo do somatório do Exemplo 7.2.

```

1 # Entrando com os dados no R
2 x <- c(50.5, 64.3, 72.6)
3
4 # Mostrando os dados armazenados
5 x
6
7 # Somatório de x
8 sum(x)

```

Lista 7.1: Comandos do software R

Na notação de somatório a variação do índice i pode não ir de 1 a n mas estar em qualquer subintervalo desses limites como pode ser observado no Exemplo 7.3.

■ **Exemplo 7.3** Seja a amostra $X = \{6, 5, 2, 8, 3, 7\}$ de tamanho $n = 6$. Determine:

a) $\sum_{i=1}^n x_i$ b) $\sum_{i=2}^5 x_i$.

Solução

a) $\sum_{i=1}^n x_i = \sum_{i=1}^6 x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 6 + 5 + 2 + 8 + 3 + 7 = 31$.

b) $\sum_{i=2}^5 x_i = x_2 + x_3 + x_4 + x_5 = 5 + 2 + 8 + 3 = 18.$

■

7.2.3 Propriedades de somatório

- a) $\sum_{i=1}^n ax_i = ax_1 + ax_2 + \dots + ax_n = a \sum_{i=1}^n x_i$
- b) $\sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + \dots + x_n y_n \neq \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$
- c) $\sum_{i=1}^n (ax_i + by_i) = ax_1 + by_1 + ax_2 + by_2 + \dots + ax_n + by_n = a \sum_{i=1}^n x_i + b \sum_{i=1}^n y_i$
- d) $\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \dots + x_n^2 \neq \left(\sum_{i=1}^n x_i \right)^2$
- e) $\sum_{i=1}^n k = \underbrace{k + k + \dots + k}_{n \text{ vezes}} = nk$

em que a, b e k são constantes.

Utilize as propriedades de somatório apresentadas anteriormente para fazer o Exercício 7.1.

Exercício 7.1 Sejam as amostras de tamanho $n = 5$ dadas por

$$X = \{2, 7, 4, 3, 2\} \text{ e } Y = \{1, 2, 3, 6, 5\}.$$

Determine:

- a) $\sum_{i=1}^n x_i$ b) $\sum_{i=1}^n y_i$ c) $\sum_{i=1}^n x_i y_i$ d) $\sum_{i=1}^n (3x_i + 2y_i)$ e) $\sum_{i=1}^n x_i^2$
- f) $\left(\sum_{i=1}^n x_i \right)^2$ g) $\sum_{i=1}^n y_i^2 + \left(\sum_{i=1}^n y_i \right)^2$

■

Na Lista 7.2 são apresentados os comandos do software R para fazer os cálculos dos somatórios solicitados no Exercício 7.1. Use as saídas do R como gabarito para este exercício.

```

1 # Entrando com os dados no R
2 x <- c(2, 7, 4, 3, 2)
3 Y <- c(1, 2, 3, 6, 5)
4
5 # Item (a)
6 sum(x)
7
8 # Item (b)
9 sum(y)
10
11 # Item (c)
12 sum(x*y)

```

```

13 # Item (d)
14 sum(3*x+2*y)
15
16 # Item (e)
17 sum(x^2)
18
19 # Item (f)
20 (sum(x) )^2
21
22 # Item (g)
23 sum(y^2) + (sum(y) )^2
24

```

Lista 7.2: Comandos do software R

7.3 Média aritmética

Dentre as medidas de posição a mais utilizada é a média. Ela é resultante da soma de todas as observações dividida pelo total de observações.

7.3.1 Média aritmética (para dados brutos)

Se x_1, x_2, \dots, x_n são n valores de uma variável quantitativa X , então a média aritmética (ou simplesmente média) de X é dada pela Equação 7.2.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (7.2)$$

Lê-se "x barra" é igual ao somatório de x dividido por n .

■ **Exemplo 7.4** Considere os dados brutos dos salários (em x sal. mín.) de 36 indivíduos:

5,73	13,60	13,23	8,46	17,26	16,22	8,74	23,30	7,39
11,06	13,85	8,12	15,99	10,76	6,26	9,80	5,25	9,77
19,40	10,53	11,59	14,69	8,95	9,35	4,56	4,00	9,13
14,71	12,00	7,59	7,44	6,66	12,79	18,75	6,86	16,61

A média dos salários (ou salário médio) é calculada por:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5,73 + 13,60 + \dots + 16,61}{36} = \frac{400,4}{36} = 11,12 \text{ salários mínimos.}$$

■

Na Lista 7.3 são apresentados os comandos do software R para calcular a média do Exemplo 7.4. O comando do R usado para calcular a média é o comando `mean()`.

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3           11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4           19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5           14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)

```

```

6
7   # Média
8   mean(dados)

```

Lista 7.3: Comandos do software R

7.3.2 Média aritmética (para dados agrupados)

Quando os dados são discretos e em grande número, pode haver repetição de valores. Nesses casos, é razoável agrupar os dados em uma tabela de distribuição de frequências. Veja a Tabela 7.1.

Tabela 7.1: Distribuição de frequências para dados discretos

Valores da variável (x_i)	Frequências (f_i)
x_1	f_1
x_2	f_2
\vdots	\vdots
x_k	f_k
Total	$\sum f_i$

A média aritmética de dados agrupados em uma tabela de distribuição de frequências é dada pela Equação 7.3.

$$\bar{x} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i} \quad (7.3)$$

Observação: Note que $x_i f_i = \underbrace{x_i + x_i + \dots + x_i}_{f_i \text{ vezes}}$, então $\sum_{i=1}^n x_i f_i = x_1 f_1 + x_2 f_2 + \dots + x_k f_k$ é igual ao somatório de todos os dados, e, $\sum_{i=1}^n f_i = f_1 + f_2 + \dots + f_k = n$, ou seja, a soma de todas as frequências é igual ao número de elementos do conjunto de dados.

■ **Exemplo 7.5** Considere a tabela de distribuição de frequências do número de filhos de vinte funcionários (Tabela 7.2).

Tabela 7.2: Distribuição de frequências do número de filhos de vinte funcionários

Número de filhos (x_i)	Frequências (f_i)	Produtos ($x_i f_i$)
0	6	$0 \times 6 = 0$
1	8	$1 \times 8 = 8$
2	4	$2 \times 4 = 8$
3	1	$3 \times 1 = 3$
4	0	$4 \times 0 = 0$
5	1	$5 \times 1 = 5$
Total	$\sum f_i = 20$	$\sum x_i f_i = 24$

A média do número de filhos é calculada por:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{24}{20} = 1,2 \text{ filhos.}$$

Na Lista 7.4 são apresentados os comandos do software R para calcular o número médio de filhos (Exemplo 7.5). Um comando do R que pode ser usado para calcular a média de dados agrupados é o mesmo comando usado para calcular a média ponderada, ou seja, o comando `weighted.mean()`.

```

1 # Entrando com os dados (xi) e as frequências (fi)
2 xi <- c(0, 1, 2, 3, 4, 5)
3 fi <- c(6, 8, 4, 1, 0, 1)
4
5 # Média dos dados agrupados
6 weighted.mean(xi, fi)
```

Lista 7.4: Comandos do software R

Quando os dados são contínuos e estão agrupados em uma tabela de distribuição de frequências a média também pode ser obtida pela Equação 7.3 com a diferença que o x_i , neste caso, representa o ponto médio da classe i . Assim, para calcular a média para dados agrupados em classes deve-se, primeiro, calcular os pontos médios das classes:

$$x_i = \frac{LI_i + LS_i}{2}$$

em que LI_i é o limite inferior da classe i e LS_i é o limite superior da classe i . Por exemplo, o ponto médio da classe $4 \leftarrow 8$ é calculado por:

$$x_i = \frac{4 + 8}{2} = 6.$$

■ **Exemplo 7.6** Considere a distribuição de frequências dos salários de 36 indivíduos apresentada na Tabela 5.5. Para facilitar o cálculo da média dos salários, a partir desta distribuição de frequências, pode-se utilizar uma tabela de cálculos auxiliares (Tabela 7.3).

Tabela 7.3: Tabela de cálculos auxiliares para obtenção da média

Classes de salários	Pontos médios (x_i)	Frequências (f_i)	Produtos ($x_i f_i$)
4,00 \leftarrow 8,00	6	10	$6 \times 10 = 60$
8,00 \leftarrow 12,00	10	12	$10 \times 12 = 120$
12,00 \leftarrow 16,00	14	8	$14 \times 8 = 112$
16,00 \leftarrow 20,00	18	5	$18 \times 5 = 90$
20,00 \leftarrow 24,00	22	1	$22 \times 1 = 22$
Total	—	$\sum f_i = 36$	$\sum x_i f_i = 404$

Utilizando os somatórios obtidos na Tabela 7.3 calcula-se a média dos salários:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{404}{36} = 11,22 \text{ salários mínimos.}$$

Observação: No Exemplo 7.4 o resultado da média calculada a partir dos dados brutos dos salários foi $\bar{x} = 11,12$. Já no Exemplo 7.6 em que foram utilizados os dados agrupados dos salários o resultado foi $\bar{x} = 11,22$. Essa diferença nos valores das duas médias ocorreu porque quando os dados foram agrupados em classes (Tabela 7.3) perdeu-se informação sobre os dados. Pela Tabela 7.3 pode-se ver, por exemplo, que a primeira classe: 4,00 – 8,00, tem frequência igual a 10, ou seja, sabe-se que 10 valores do conjunto de dados estão entre 4,00 e 8,00, porém, não se sabe (olhando apenas nesta tabela) quais são estes valores. Assim, quando utiliza-se o ponto médio ($x_i = 6$) para representar esta classe no cálculo da média assume-se que todos os 10 valores dentro desta classe são iguais a 6, ou seja, usa-se uma aproximação para os valores desta classe. Logo, a média calculada a partir de dados agrupados em classes é uma média aproximada. Como a média calculada a partir dos dados brutos é mais precisa (pois nenhuma informação sobre os dados foi perdida) sempre que se tiver acesso aos dados brutos eles devem ser preferidos para efetuar o cálculo da média.

Na Lista 7.5 são apresentados os comandos do software R para calcular o salário médio (Exemplo 7.6).

```

1 # Entrando com os pontos médios (xi) e as frequências (fi)
2 xi <- c(6, 10, 14, 18, 22)
3 fi <- c(10, 12, 8, 5, 1)
4
5 # Média dos dados agrupados
6 weighted.mean(xi, fi)
```

Lista 7.5: Comandos do software R

Propriedades da média

1. é única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado;
2. por depender de todos os valores observados, qualquer modificação nos dados fará com que a média fique alterada;
3. é afetada por valores atípicos observados, o que a torna uma medida inadequada para representar variáveis com a presença desses valores;

O que é um valor atípico da variável?

Valor que destoa em magnitude dos demais valores do conjunto estudado. Também é denominado de *outlier*.

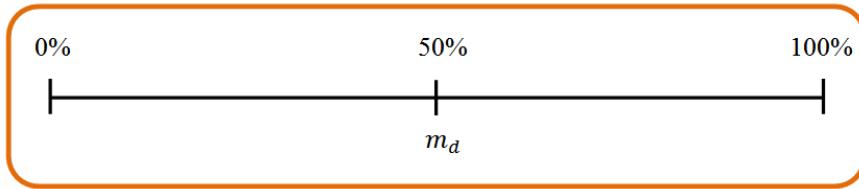
4. a soma da diferença de cada valor observado em relação à média é zero, ou seja, a soma dos desvios é zero:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0;$$

5. a soma dos quadrados dos desvios tomados em relação à média aritmética é um mínimo. Qualquer valor que não seja a média aritmética resultará em um valor superior a $\sum_{i=1}^n (x_i - \bar{x})^2$;
6. somando ou subtraindo uma constante não nula aos valores da distribuição da variável, a média aritmética "receberá" a soma ou subtração da constante.
7. multiplicando ou dividindo uma constante não nula aos valores da variável, a média ficará multiplicada ou dividida pela constante;

7.4 Mediana

A mediana (m_d) é o valor que divide o conjunto de dados ordenados em duas partes iguais:



7.4.1 Mediana para dados brutos

Número ímpar de dados (n ímpar)

Quando o número de dados é ímpar existe um único elemento na posição central do conjunto de dados ordenados. Neste caso, a mediana é igual a este elemento.

■ **Exemplo 7.7** O conjunto de dados ordenados

$$\overline{1 \quad 3 \quad 5 \quad 7 \quad 9}$$

tem mediana $m_d = 5$ pois este é o elemento central deste conjunto de dados ordenados. Observe que este valor divide o conjunto de dados ordenados em duas partes iguais.

■

Número par de dados (n par)

Quando o número de dados é par existem dois elementos centrais no conjunto de dados ordenados. Neste caso, a mediana é igual a média desses dois valores.

■ **Exemplo 7.8** O conjunto de dados ordenados

$$\overline{3 \quad 5 \quad 7 \quad 9}$$

tem mediana $m_d = 6$ pois este valor é a média dos dois elementos centrais do conjunto de dados ordenados, ou seja,

$$m_d = \frac{5+7}{2} = 6.$$

Observe que este valor divide o conjunto de dados ordenados em duas partes iguais. Note também que a mediana não precisa ser igual a nenhum dos elementos do conjunto de dados.

■

Fórmula para calcular a mediana

Pode-se, ainda, utilizar a fórmula apresentada na Equação 7.4 para calcular a mediana.

$$m_d = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ for ímpar;} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{se } n \text{ for par.} \end{cases} \quad (7.4)$$

■ **Exemplo 7.9** Considere o conjunto de dados ordenados:

1	3	5	7	9
---	---	---	---	---

Como $n = 5$ é ímpar, então:

$$m_d = x_{(\frac{n+1}{2})} = x_{(\frac{5+1}{2})} = x_{(\frac{6}{2})} = x_3 = 5.$$

Considere, agora, o conjunto de dados ordenados:

3	5	7	9
---	---	---	---

Como $n = 4$ é par, então:

$$m_d = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} = \frac{x_{(\frac{4}{2})} + x_{(\frac{4}{2}+1)}}{2} = \frac{x_2 + x_3}{2} = \frac{5 + 7}{2} = \frac{12}{2} = 6.$$

■

Na Lista 7.6 são apresentados os comandos do software R para calcular a mediana.

```

1 # Entrando com os dados no R (n ímpar)
2 dados1 <- c(1, 3, 5, 7, 9)
3
4 # Mediana
5 median(dados1)
6
7 # Entrando com os dados no R (n par)
8 dados2 <- c(3, 5, 7, 9)
9
10 # Mediana
11 median(dados2)

```

Lista 7.6: Comandos do software R

7.4.2 Mediana para dados agrupados

Mediana para dados discretos agrupados

Quando dados quantitativos discretos estão agrupados em uma tabela de distribuição de frequências, utiliza-se a frequência acumulada para verificar em qual classe está o elemento da posição central $x_{(\frac{n+1}{2})}$, caso n for ímpar, ou os dois elementos centrais $x_{(\frac{n}{2})}$ e $x_{(\frac{n}{2}+1)}$, caso n for par. A frequência acumulada de uma determinada linha é a soma das frequências das linhas anteriores com a frequência desta linha (Tabela 7.4).

Tabela 7.4: Obtenção das frequências acumuladas

Dados (x_i)	Frequências (f_i)	Frequências acumuladas (F_{ac})
x_1	f_1	f_1
x_2	f_2	$f_1 + f_2$
x_3	f_3	$f_1 + f_2 + f_3$
\vdots	\vdots	\vdots
x_k	f_k	$f_1 + f_2 + f_3 + \dots + f_k$
Total	$\sum f_i$	—

■ **Exemplo 7.10** Considere a tabela de distribuição de frequências do número de filhos de vinte funcionários, contendo as frequências acumuladas (Tabela 7.5).

Tabela 7.5: Distribuição de frequências do número de filhos de vinte funcionários

Número de filhos (x_i)	Frequências (f_i)	Frequências acumuladas (F_{ac})
0	6	6
1	8	14
2	4	18
3	1	19
4	0	19
5	1	20
Total	$\sum f_i = 20$	—

Como $n = 20$ (n par), então a mediana é a média dos dois elementos centrais: x_{10} e x_{11} . Observando as frequências acumuladas pode-se ver que x_{10} e x_{11} não estão na primeira linha da tabela, pois, a frequência acumulada da primeira linha é 6, ou seja, até a primeira linha tem apenas 6 elementos (vai até o elemento x_6). Como a frequência acumulada da segunda linha é 14 então até a segunda linha tem 14 elementos (vai desde o elemento x_7 até o elemento x_{14}), ou seja, os elementos x_{10} e x_{11} estão ambos na segunda linha. Como na segunda linha todos os elementos são iguais ao número 1 (isto é, $x_i = 1$), então:

$$m_d = \frac{x_{10} + x_{11}}{2} = \frac{1 + 1}{2} = 1.$$

■

Na Lista 7.7 são apresentados os comandos do software R para calcular a mediana para dados discretos agrupados em uma distribuição de frequências.

```

1 # Entrando com os dados (xi) e as frequências (fi) no R
2 xi <- c(0, 1, 2, 3, 4, 5)
3 fi <- c(6, 8, 4, 1, 0, 1)
4
5 # Desagrupando os dados no R
6 dados <- rep(xi, fi); dados
7
8 # Mediana
9 median(dados)
```

Lista 7.7: Comandos do software R

Mediana para dados contínuos agrupados

Para dados quantitativos contínuos agrupados em uma tabela de distribuição de frequências (em classes) deve-se determinar a classe mediana, que é a classe que contém o elemento da posição $n/2$ (independentemente de n ser par ou ímpar). Depois, determina-se o valor da mediana utilizando a Equação 7.5.

$$m_d = LI_{md} + \frac{\frac{n}{2} - F_{ac}^*}{f_{md}} \times c_{md} \quad (7.5)$$

em que:

- LI_{md} é o limite inferior da classe mediana;
- c_{md} é a amplitude da classe mediana ($c_{md} = LS_{md} - LI_{md}$);
- f_{md} é a frequência da classe mediana;
- F_{ac}^* é a frequência acumulada da **classe anterior** à classe mediana. Se a classe mediana for a primeira classe então F_{ac}^* será igual a zero.

■ **Exemplo 7.11** Na Tabela 7.6 é apresentada a distribuição de frequências dos salários de 100 indivíduos (em x sal. mín.).

Tabela 7.6: Distribuição de frequências dos salários de 100 indivíduos

Classes	Frequências (f_i)	Frequências acumuladas (F_{ac})
1,5 ⊢ 2,0	3	3
2,0 ⊢ 2,5	16	19
2,5 ⊢ 3,0	32	51
3,0 ⊢ 3,5	33	84
3,5 ⊢ 4,0	11	95
4,0 ⊢ 4,5	4	99
4,5 ⊢ 5,0	1	100
Total	100	—

Como $n = 100$, então $n/2 = 100/2 = 50$, e dessa forma, o elemento x_{50} está na terceira classe pois a frequência acumulada da terceira classe é igual a 51 (esta classe inclui os elementos de x_{20} até x_{51}). Então a terceira classe é a classe mediana (classe que contém a mediana).

Na Figura 7.1 são apresentadas as localizações na tabela dos itens necessários para o cálculo da mediana.

Assim, os itens necessários para efetuar o cálculo da mediana são:

- $n/2 = 100/2 = 50$;
- $LI_{md} = 2,5$;
- $F_{ac}^* = 19$;
- $f_{md} = 32$;
- $c_{md} = 3,0 - 2,5 = 0,5$.

Classes	Frequências (f_i)	Frequências acumuladas (F_{ac})
1,5 ⊢ 2,0	3	3
2,0 ⊢ 2,5	16	19 F_{ac}^*
LI_{md} 2,5 ⊢ 3,0	32 f_{md}	51 Contém $x_{n/2}$
3,0 ⊢ 3,5	33	84
3,5 ⊢ 4,0	11	95
4,0 ⊢ 4,5	4	99
4,5 ⊢ 5,0	1	100
Total	100	—

Figura 7.1: Itens para o cálculo da mediana

Portanto, a mediana é calculada por:

$$m_d = LI_{md} + \frac{\frac{n}{2} - F_{ac}}{f_{md}} \times c_{md}$$

$$= 2,5 + \frac{50 - 19}{32} \times 0,5$$

$$= 2,985 \text{ salários.}$$

■

Propriedades da mediana

- é única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado;
- não depende de todos os valores do conjunto de dados, podendo não se alterar com a modificação de alguns deles;
- não é influenciada por valores atípicos do conjunto de dados;
- somando ou subtraindo uma constante não nula aos valores da variável a mediana "receberá" a soma ou subtração da constante;
- multiplicando ou dividindo uma constante não nula aos valores da variável, a mediana ficará multiplicada ou dividida pela constante.

7.5 Moda

7.5.1 Moda para dados brutos

A moda é o elemento do conjunto de dados que ocorre com maior frequência.

■ **Exemplo 7.12** Considere o conjunto de dados :

0	0	2	5	3	7	4	7	8	7	9	6
---	---	---	---	---	---	---	---	---	---	---	---

A moda deste conjunto de dados é o número 7 porque este é o elemento que ocorre o maior número de vezes.

■

Observação: Um conjunto de dados pode não ter moda ou ter duas ou mais modas.

■ **Exemplo 7.13** Considere o conjunto de dados :

0	2	4	6	8	10
---	---	---	---	---	----

Este conjunto de dados não tem moda pois nenhum elemento ocorre com maior frequência.

Considere, agora, o conjunto de dados:

1	2	2	3	4	4	5	6	7
---	---	---	---	---	---	---	---	---

Este conjunto de dados tem duas modas: 2 e 4.

■

Na Lista 7.9 são apresentados os comandos do software R para calcular a moda para dados brutos.

```

1 # Entrando com os dados no R
2 dados <- c(0, 0, 2, 5, 3, 7, 4, 7, 8, 7, 9, 6)
3
4 # Contando as frequências dos elementos
5 tab <- table(dados)
6
7 # Moda (elemento com maior frequência)
8 as.numeric(names(tab)[tab==max(tab)])

```

Lista 7.8: Comandos do software R

7.5.2 Moda para dados agrupados

Moda para dados discretos agrupados

Para dados quantitativos discretos agrupados em uma tabela de distribuição de frequências a moda é o elemento que tem a maior frequência (elemento que ocorre maior número de vezes).

■ **Exemplo 7.14** Considere os dados quantitativos discretos agrupados em uma distribuição de frequências (Tabela 7.7).

A moda é o número 7 pois $x_i = 7$ é o valor que tem a maior frequência ($f_i = 3$). Portanto, $m_o = 7$.

■

Na Lista 7.9 são apresentados os comandos do software R para calcular a moda para dados discretos agrupados em uma distribuição de frequências.

Tabela 7.7: Distribuição de frequências de dados quantitativos discretos

Dados (x_i)	Frequências (f_i)
0	2
1	0
2	1
3	1
4	1
5	1
6	1
7	3
8	1
9	1
Total	12

```

1 # Entrando com os dados no R
2 xi <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
3 fi <- c(2, 0, 1, 1, 1, 1, 1, 3, 1, 1)
4
5 # Moda (elemento com maior frequência)
6 xi[fi==max(fi)]

```

Lista 7.9: Comandos do software R

Moda para dados contínuos agrupados

Para dados quantitativos contínuos a moda é o valor de maior densidade, ou seja, é o valor sob o qual a variável atinge o ponto mais alto de sua distribuição.

Para dados contínuos agrupados em uma tabela de distribuição de frequências (em classes), a classe modal (classe que contém a moda) é a classe com a maior frequência. Para determinar o valor da moda utiliza-se a Equação 7.6 (fórmula de Czuber).

$$m_o = LI_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c_{mo} \quad (7.6)$$

em que:

- LI_{mo} é o limite inferior da classe modal;
- c_{mo} é a amplitude da classe modal;
- Δ_1 é a diferença entre a frequência da classe modal e a frequência da classe imediatamente anterior;
- Δ_2 é a diferença entre a frequência da classe modal e a frequência da classe imediatamente posterior.

■ **Exemplo 7.15** Na Tabela 7.8 é apresentada a distribuição de frequências dos salários de 100 indivíduos (em x sal. mín.).

Tabela 7.8: Distribuição de frequências dos salários de 100 indivíduos

Classes	Frequências (f_i)
1,5 ⊢ 2,0	3
2,0 ⊢ 2,5	16
2,5 ⊢ 3,0	32
3,0 ⊢ 3,5	33
3,5 ⊢ 4,0	11
4,0 ⊢ 4,5	4
4,5 ⊢ 5,0	1
Total	100

Observe que a classe com maior frequência é a quarta classe (frequência igual a 33). Portanto a quarta classe é a classe modal.

Os itens necessários para efetuar o cálculo da moda são:

- $LI_{mo} = 3,0$;
- $\Delta_1 = 33 - 32 = 1$;
- $\Delta_2 = 33 - 11 = 22$;
- $c_{mo} = 3,5 - 3,0 = 0,5$.

Na Figura 7.2 são apresentadas as localizações na tabela dos itens necessários para o cálculo da moda.

Classes	Frequências (f_i)
1,5 ⊢ 2,0	3
2,0 ⊢ 2,5	16
2,5 ⊢ 3,0	32
3,0 ⊢ 3,5	33
3,5 ⊢ 4,0	11
4,0 ⊢ 4,5	4
4,5 ⊢ 5,0	1
Total	100

Figura 7.2: Itens para o cálculo da moda

Portanto, a moda é calculada por:

$$m_o = LI_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c_{mo}$$

$$= 3,0 + \frac{1}{1+22} \times 0,5$$

$$= 3,022 \text{ salários.}$$

■

7.5.3 Moda para dados qualitativos

A moda é a única medida de posição que também pode ser usada para descrever dados qualitativos. Nesse caso, a moda é a categoria da variável que ocorre com maior frequência.

■ **Exemplo 7.16** Na Tabela 7.9 é apresentada a distribuição de frequências do tipo sanguíneo de 1.167 indivíduos.

Tabela 7.9: Distribuição de frequências do tipo sanguíneo de 1.167 indivíduos

Grupo sanguíneo	Frequências (f_i)
O	550
A	456
B	132
AB	29
Total	1167

A moda é o grupo sanguíneo O pois esta foi a categoria que ocorreu com maior frequência ($f_i = 550$). ■

Propriedades da moda

- sempre é representada por um dos valores da variável;
- não depende de todos os valores do conjunto de dados, podendo não se alterar com a modificação de alguns deles;
- não é influenciada pelos valores atípicos do conjunto de dados;
- somando ou subtraindo uma constante não nula aos valores da variável, a moda aritmética "receberá" a soma ou subtração da constante;
- multiplicando ou dividindo uma constante não nula aos valores da variável, a moda ficará multiplicada ou dividida pela constante.

7.6 Utilização das medidas de tendência central

Costa (2012) propõe os seguintes critérios para escolher entre as medidas de posição:

Escolha da média:

1. quando a distribuição dos dados é pelo menos aproximadamente simétrica;
2. quando for necessário obter posteriormente outros parâmetros que podem depender da média, como por exemplo a variância, o desvio padrão, etc.

Escolha da mediana:

1. quando há valores extremos;
2. quando deseja-se conhecer o ponto central da distribuição;
3. quando a distribuição dos dados é muito assimétrica.

Escolha da moda:

1. quando a medida de interesse é o ponto mais típico ou popular dos dados;
2. quando precisa-se apenas de uma rápida idéia sobre a tendência central dos dados.

7.7 Medidas separatrizes (quantis)

Foi visto que a mediana é um valor que divide o conjunto de dados ao meio, ou seja, a mediana deixa 50% dos dados abaixo dela e 50% acima. De um modo geral podemos definir uma medida chamada quantil de ordem $q(p)$, em que p é uma proporção qualquer, $0 < p < 1$, tal que $100p\%$ das observações sejam menores do que $q(p)$.

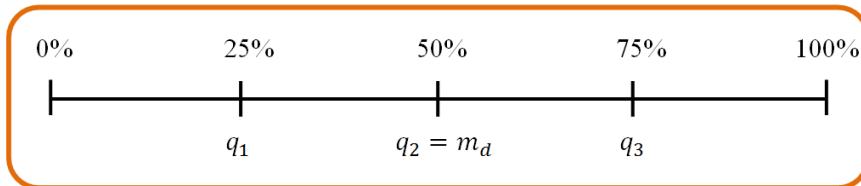
■ **Exemplo 7.17** A seguir são apresentados alguns quantis e seus nomes particulares:

- $q(0,25)$: 25º percentil \Rightarrow 25% das observações são menores do que $q(0,25)$. O 25º percentil também é conhecido como 1º quartil;
- $q(0,50)$: 50º percentil \Rightarrow 50% das observações são menores do que $q(0,50)$. Note que o 50º percentil é igual à mediana, ou seja, $q(0,50) = m_d$. A mediana também é conhecida como 2º quartil;
- $q(0,90)$: 90º percentil \Rightarrow 90% das observações são menores do que $q(0,90)$. O 90º percentil também é conhecido como 9º decil.

■

7.7.1 Quartis para dados brutos

Os quartis (q_1 , q_2 e q_3) são valores que dividem o conjunto de dados em quatro partes iguais:



O primeiro quartil (q_1) é um valor tal que 25% dos dados são menores ou iguais a ele (então os outros 75% dos dados são maiores ou iguais a ele), o segundo quartil (q_2), que é igual à mediana (m_d), é um valor tal que 50% dos dados são menores ou iguais a ele (então os outros 50% dos dados são maiores ou iguais a ele) e o terceiro quartil (q_3) é um valor tal que 75% dos dados são menores ou iguais a ele (então os outros 25% dos dados são maiores ou iguais a ele).

Para calcular os quartis primeiro deve-se ordenar o conjunto de dados.

Obtendo os quartis de um conjunto com um número ímpar de dados

■ **Exemplo 7.18** Considere o conjunto de dados:

3	4	1	5	7	9	2	10	6
---	---	---	---	---	---	---	----	---

Primeiramente deve-se ordenar o conjunto de dados:

1	2	3	4	5	6	7	9	10
---	---	---	---	---	---	---	---	----

Observe que $n = 9$ (n ímpar). Então, a mediana é o valor central dos dados ordenados, ou seja, $m_d = 5$:

1	2	3	4	5	6	7	9	10
---	---	---	---	---	---	---	---	----

Como $q_2 = m_d$ então o segundo quartil já foi obtido, ou seja, $q_2 = 5$.

Para obter o primeiro quartil (q_1) e o terceiro quartil (q_3) separe o conjunto de dados ordenados em dois subconjuntos:

- a) dados ordenados à esquerda da mediana, **incluindo a mediana**;
- b) dados ordenados à direita da mediana, **incluindo a mediana**,

e obtenha as medianas destes subconjuntos. Estas medianas serão, respectivamente, q_1 e q_3 .

Assim, para o subconjunto (a):

1	2	3	4	5
---	---	----------	---	---

a mediana é igual a 3 e, portanto, $q_1 = 3$.

Para o subconjunto (b):

5	6	7	9	10
---	---	----------	---	----

a mediana é igual a 7 e, portanto, $q_3 = 7$.

Portanto, os quartis são: $q_1 = 3$, $q_2 = 5$ e $q_3 = 7$.

1	2	3	4	5	6	7	9	10
---	---	----------	---	----------	---	---	----------	----

■

Observação: Note no Exemplo 7.18 que no conjunto de dados ordenados dois valores são menores do que q_1 e seis valores são maiores do que q_1 , quatro valores são maiores do que q_2 e quatro valores são maiores do que q_3 e, por fim, seis valores são menores do que q_3 e dois valores são maiores do que q_3 . Portanto os quartis dividiram o conjunto de dados em quatro partes aproximadamente iguais.

Na Lista 7.10 são apresentados os comandos do software R para calcular os quartis para um conjunto com um número ímpar de dados.

```

1 # Entrando com os dados no R
2 dados <- c(3, 4, 1, 5, 7, 9, 2, 10, 6)
3
4 # Número de elementos do conjunto de dados
5 length(dados)
6
7 # Quartis (para n ímpar)
8 quantile(dados)
9
10 # Quartis e outras medidas (mínimo, máximo e média)
11 summary(dados)

```

Lista 7.10: Comandos do software R

Obtendo os quartis de um conjunto com um número par de dados

■ **Exemplo 7.19** Considere o conjunto de dados:

11	3	4	1	5	7	9	2	10	6
----	---	---	---	---	---	---	---	----	---

Primeiramente deve-se ordenar o conjunto de dados:

1	2	3	4	5	6	7	9	10	11
---	---	---	---	---	---	---	---	----	----

Observe que $n = 10$ (n par). Então, a mediana é a média dos dois valores centrais dos dados ordenados, ou seja,

$$m_d = \frac{5+6}{2} = 5,5.$$

Como $q_2 = m_d$ então o segundo quartil já foi obtido, ou seja, $q_2 = 5,5$. Observe que, neste caso, a mediana não é um valor pertencente ao conjunto de dados.

Para obter o primeiro quartil (q_1) e o terceiro quartil (q_3) separe o conjunto de dados ordenados em dois subconjuntos:

- a) dados ordenados à esquerda da mediana, **sem incluir a mediana**;
- b) dados ordenados à direita da mediana, **sem incluir a mediana**,

e obtenha as medianas destes subconjuntos. Estas medianas serão, respectivamente, q_1 e q_3 .

Assim, para o subconjunto (a):

1	2	3	4	5
---	---	---	---	---

a mediana é igual a 3 e, portanto, $q_1 = 3$.

Para o subconjunto (b):

6	7	9	10	11
---	---	---	----	----

a mediana é igual a 9 e, portanto, $q_3 = 9$.

Portanto, os quartis são: $q_1 = 3$, $q_2 = 5,5$ e $q_3 = 9$.

Observação: As formas de se calcular os quartis não são as únicas, existem várias formas diferentes de se calcular quartis. Na Lista 7.11 são apresentados os comandos do software R para calcular os quartis para um conjunto com um número par de dados, como foi visto no Exemplo 7.19, e também por outros métodos que não foram abordados neste material (os resultados são diferentes).

```

1 # Entrando com os dados no R
2 dados <- c(11, 3, 4, 1, 5, 7, 9, 2, 10, 6)
3
4 # Número de elementos do conjunto de dados
5 length(dados)
6

```

```

7 # Quartis (para n par, método abordado no material)
8 quantile(dados, type=5)
9
10 # Quartis (método diferente do abordado no material)
11 quantile(dados)
12
13 # Quartis e outras medidas (método diferente do abordado no material)
14 summary(dados)

```

Lista 7.11: Comandos do software R

Observação: Na Lista 7.11 o argumento `type=5` do comando `quantile()` é que seleciona o método de cálculo dos quartis, para n par, igual ao que foi usado neste material.

7.7.2 Quartis para dados agrupados

Foi visto anteriormente que a mediana para dados contínuos agrupados em classes era calculada pela Equação 7.5, ou seja,

$$m_d = LI_{md} + \frac{\frac{n}{2} - F_{ac}^*}{f_{md}} \times c_{md}.$$

Como a mediana é o segundo quartil ($m_d = q_2$), então, de maneira análoga são obtidas as fórmulas para o 1º quartil e para o 3º quartil, apresentadas a seguir.

Determinação do 1º quartil

1. calcule $n/4$;
2. identifique a classe q_1 pela frequência acumulada. A classe q_1 é a classe que contém o elemento da posição $n/4$;
3. aplique a Equação 7.7 para obter o 1º quartil:

$$q_1 = LI_{q_1} + \frac{\frac{n}{4} - F_{ac}^*}{f_{q_1}} \times c_{q_1} \quad (7.7)$$

em que:

- LI_{q_1} é o limite inferior da classe do 1º quartil;
- c_{q_1} é a amplitude da classe do 1º quartil;
- f_{q_1} é a frequência da classe do 1º quartil;
- F_{ac}^* é a frequência acumulada da **classe anterior** à classe do 1º quartil. Se a classe do 1º quartil for a primeira classe então F_{ac}^* será igual a zero.

Determinação do 3º quartil

1. calcule $3n/4$;
2. identifique a classe q_3 pela frequência acumulada. A classe q_3 é a classe que contém o elemento da posição $3n/4$;
3. aplique a Equação 7.8 para obter o 3º quartil:

$$q_3 = LI_{q_3} + \frac{\frac{3n}{4} - F_{ac}^*}{f_{q_3}} \times c_{q_3} \quad (7.8)$$

em que:

- LI_{q_3} é o limite inferior da classe do 3º quartil;
- c_{q_3} é a amplitude da classe do 3º quartil;
- f_{q_3} é a frequência da classe do 3º quartil;
- F_{ac}^* é a frequência acumulada da **classe anterior** à classe do 3º quartil. Se a classe do 3º quartil for a primeira classe então F_{ac}^* será igual a zero.

■ **Exemplo 7.20** Considere a distribuição de frequências dos pesos, em kg, de 56 crianças e adolescentes apresentada na Tabel 7.10.

Tabela 7.10: Distribuição de frequências dos pesos, em kg, de 56 crianças e adolescentes

Pesos (kg)	Frequências (f_i)	Frequências acumuladas (F_{ac})	
7 ⊢ 17	6	6	
17 ⊢ 27	15	21	⇐ classe q_1
27 ⊢ 37	20	41	⇐ classe q_2
37 ⊢ 47	10	51	⇐ classe q_3
47 ⊢ 57	5	56	
Total	56	—	

Observando as frequências acumuladas são obtidas as classes que contém os quartis (classe q_1 , classe q_2 e classe q_3):

- $\frac{n}{4} = \frac{56}{4} = 14 \Rightarrow$ a 2ª classe ($F_{ac} = 21 \Rightarrow$ elementos: x_7 até x_{21}) contém o elemento da posição $n/4 = 14$. Portanto esta é a classe q_1 ;
- $\frac{n}{2} = \frac{56}{2} = 28 \Rightarrow$ a 3ª classe ($F_{ac} = 41 \Rightarrow$ elementos: x_{22} até x_{41}) contém o elemento da posição $n/2 = 28$. Portanto esta é a classe q_2 ;
- $\frac{3 \times n}{4} = \frac{3 \times 56}{4} = 42 \Rightarrow$ a 4ª classe ($F_{ac} = 51 \Rightarrow$ elementos: x_{42} até x_{51}) contém o elemento da posição $3n/4 = 42$. Portanto esta é a classe q_3 .

Após encontrar as classes deve-se obter os itens das fórmulas (Equações 7.7, 7.5 e 7.8):

- Itens para calcular q_1 : $LI_{q_1} = 17$, $n/4 = 14$, $F_{ac}^* = 6$, $f_{q_1} = 15$, $c_{q_1} = 10$;
- Itens para calcular $q_2 = m_d$: $LI_{m_d} = 27$, $n/2 = 28$, $F_{ac}^* = 21$, $f_{m_d} = 20$, $c_{m_d} = 10$;
- Itens para calcular q_3 : $LI_{q_3} = 37$, $3n/4 = 42$, $F_{ac}^* = 41$, $f_{q_3} = 10$, $c_{q_3} = 10$.

Após obter os itens das fórmulas basta substituí-los nas respectivas equações para calcular os quartis:

1º quartil:

$$q_1 = LI_{q_1} + \frac{\frac{n}{4} - F_{ac}^*}{f_{q_1}} \times c_{q_1}$$

$$= 17 + \frac{14 - 6}{15} \times 10$$

$$= 22,3 \text{ kg}$$

2º quartil:

$$q_2 = m_d = LI_{m_d} + \frac{\frac{n}{2} - F_{ac}^*}{f_{m_d}} \times c_{m_d}$$

$$= 27 + \frac{28 - 21}{20} \times 10$$

$$= 30,5 \text{ kg}$$

3º quartil:

$$q_3 = LI_{q_3} + \frac{\frac{3 \times n}{4} - F_{ac}^*}{f_{q_3}} \times c_{q_3}$$

$$= 37 + \frac{42 - 41}{10} \times 10$$

$$= 38,0 \text{ kg}$$

■

7.7.3 Percentis

Dependendo do quantil desejado, por exemplo, o 22º percentil (P_{22}), pode haver dificuldades em calculá-lo para dados brutos. Sendo assim, pode-se calcular os percentis utilizando dados agrupados.

O percentil de ordem i , em que i é um número de 1 a 100, pode ser calculado pela Equação 7.9.

$$P_i = LI_{P_i} + \frac{\frac{i \times n}{100} - F_{ac}^*}{f_{P_i}} \times c_{P_i} \quad (7.9)$$

em que:

- LI_{P_i} é o limite inferior da classe P_i ;
- n é o tamanho da amostra;
- F_{ac}^* é a frequência acumulada da **classe anterior** à classe P_i ;
- f_{P_i} é a frequência da classe P_i ;
- c_{P_i} é a amplitude da classe P_i .

Observação: A classe P_i é a classe que contém o elemento da posição $\frac{i \times n}{100}$. Para localizar esta classe deve-se utilizar as frequências acumuladas.

Exercício 7.2 Utilizando a distribuição de frequências dos pesos de crianças e adolescentes (Tabela 7.10), apresentada no Exemplo 7.20, determine e interprete o 90º percentil (P_{90}). ■

7.8 Boxplot

As medidas que foram abordadas em seções anteriores: mínimo, máximo e quartis, permitem traçar o *Boxplot* (ou diagrama de caixa) que ajuda a entender a informação contida em um conjunto de dados.

O *Boxplot* dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão é dada pela distância interquartílica. As posições relativas de q_1 , q_2 e q_3 no *Boxplot* também dão uma noção da assimetria (será visto no capítulo 9) da distribuição da variável.

Na Figura 7.3 são apresentados os elementos de um *Boxplot* e também uma visão geral da aparência deste tipo de gráfico.

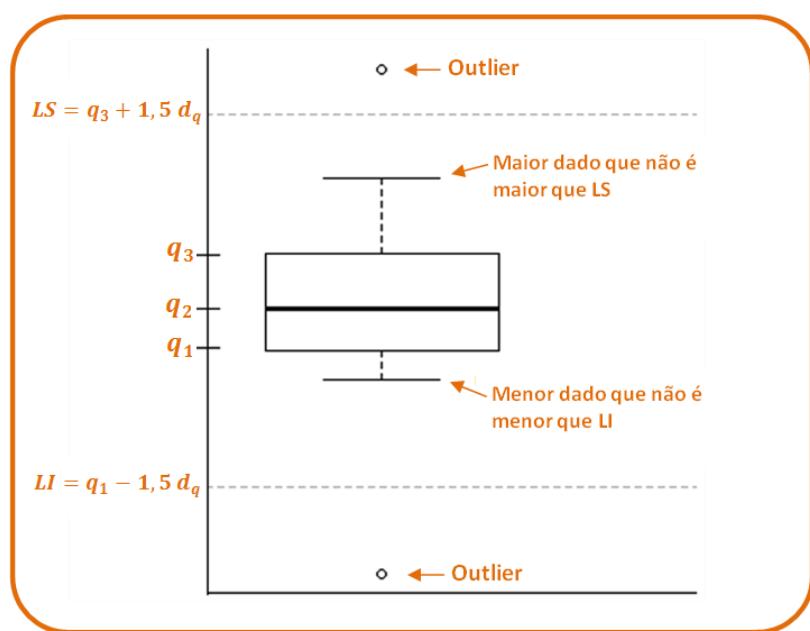


Figura 7.3: Elementos do Boxplot

Passos para construir o Boxplot

1. desenhe um segmento de reta na posição vertical, para representar a amplitude dos dados;
2. marque nesse segmento, o primeiro, o segundo e o terceiro quartis, obedecendo a escala;
3. desenhe um retângulo de maneira que o lado inferior e o lado superior passem exatamente nas alturas dos pontos que marcam o primeiro e o terceiro quartis;
4. faça um traço diagonal dentro do retângulo na altura do ponto que marca a mediana;
5. calcule um limite inferior (LI) e um limite superior (LS) da seguinte maneira:

$$LI = q_1 - 1,5 \times d_q \quad \text{e} \quad LS = q_3 + 1,5 \times d_q$$

em que $d_q = q_3 - q_1$ é a distância interquartílica;

6. a partir do retângulo, para cima, segue uma linha até o maior valor dos dados que não seja maior que LS ;
7. a partir do retângulo, para baixo, segue uma linha até o menor valor dos dados que não seja menor que LI ;
8. as observações (dados) que forem maiores ou iguais ao limite superior ou menores ou iguais ao limite inferior são valores atípicos, também chamados de *outliers*, e são representados no gráfico por "bolinhas".

■ **Exemplo 7.21** Considere o conjunto de dados ordenados:

1	2	3	4	5	6	7	10	16
---	---	----------	---	----------	---	----------	----	----

Os quartis deste conjunto de dados são:

- $q_1 = 3$;
- $q_2 = m_d = 5$;
- $q_3 = 7$.

Outros elementos necessários para construir o Boxplot:

- $d_q = q_3 - q_1 = 7 - 3 = 4$;
- $LI = q_1 - 1,5 \times d_q = 3 - 1,5 \times 4 = -3$;
- $LS = q_3 + 1,5 \times d_q = 7 + 1,5 \times 4 = 13$;
- menor valor do conjunto de dados que não é menor do que LI : $x = 1$;
- maior valor do conjunto de dados que não é maior do que LS : $x = 10$;
- *outliers* (valores menores do que LI ou maiores do que LS): $x = 16$.

Utilizando os quartis e os demais elementos apresentados acima pode-se construir o Boxplot (Figura 7.4). ■

Na Lista 7.12 são apresentados os comandos do software R para gerar o Boxplot.

```

1 # Entrando com os dados no R
2 dados <- c(1, 2, 3, 4, 5, 6, 7, 10, 16)
3
4 # Gerando o Boxplot
5 boxplot(dados, ylab="Nome da variável")
6
7 # Alterando os valores do eixo y do Boxplot
8 boxplot(dados, ylab="Nome da variável", axes=F, ylim=c(0,17))

```

```

9 axis(2,c(1,3,5,7,10,16))
10 box()

```

Lista 7.12: Comandos do software R

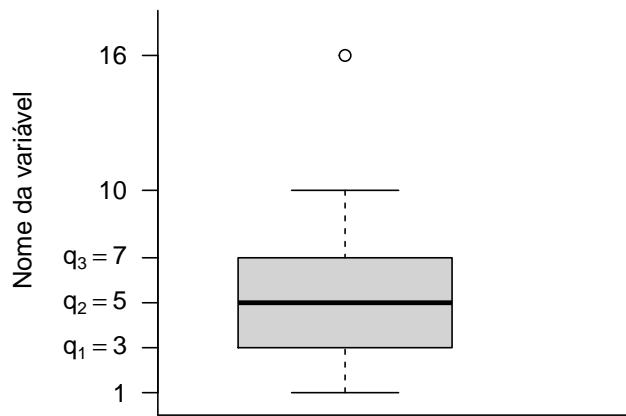


Figura 7.4: Boxplot

7.8.1 Boxplot comparativo

Outra utilidade do Boxplot é na comparação de diferentes conjuntos de dados.

■ **Exemplo 7.22** Foi feito um experimento para comparar dois métodos de treinamento para a execução de um serviço especializado. Vinte homens foram selecionados para esse treinamento. Dez foram escolhidos ao acaso e treinados pelo método A. Outros dez foram treinados pelo método B. Concluído o período de treinamento, todos os homens executaram o serviço e foi medido o tempo de cada um. Os dados são apresentados na Tabela 7.11.

Tabela 7.11: Tempo (em minutos) para execução do serviço, segundo o método de treinamento

Método A	Método B
11	23
20	46
5	13
23	19
16	23
21	17
18	28
16	36
27	25
24	28

A comparação do tempo de execução de serviço pelos métodos A e B pode ser feita utilizando um Boxplot comparativo que consiste em dois Boxplots em um mesmo sistema de eixos cartesianos. O Boxplot comparativo dos métodos de treinamento é apresentado na Figura 7.5.

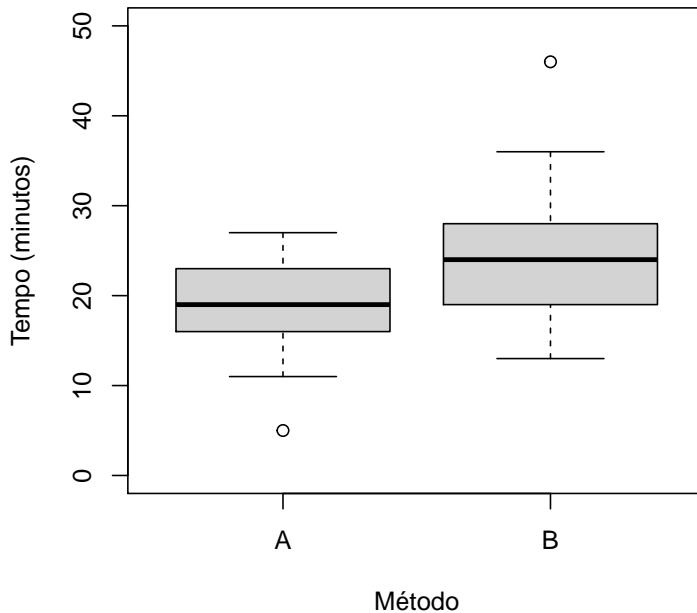


Figura 7.5: Boxplot comparativo dos métodos de treinamento

Observe pela Figura 7.5 que os homens que foram treinados pelo método A parecem estar executando o serviço em um tempo menor do que os que foram treinados pelo método B.

■
Na Lista 7.13 são apresentados os comandos do software R para gerar um Boxplot comparativo.

```

1 # Entrando com os dados no R
2 A <- c(11, 20, 5, 23, 16, 21, 18, 16, 27, 24)
3 B <- c(23, 46, 13, 19, 23, 17, 28, 36, 25, 28)
4
5 # Gerando o Boxplot comparativo
6 boxplot(A,B,names=c("A", "B"),xlab="Método",ylab="Tempo (minutos)")
7
8 # Alterando os valores do eixo y do Boxplot
9 boxplot(A,B,xlab="Método",ylab="Tempo (minutos)",axes=F,ylim=c(0,50))
10 axis(1,at=c(1,2),labels=c("A", "B"))
11 axis(2,c(0,10,20,30,40,50))
12 box()
```

Lista 7.13: Comandos do software R

Observação: Considerando ainda os dados do Exemplo 7.22, outra forma de entrada de dados no R para fazer um Boxplot comparativo é usando a mesma estrutura da Tabela 7.12.

Na Lista 7.14 são apresentados os comandos do software R para gerar um Boxplot comparativo com dados na mesma estrutura da Tabela 7.12.

```

1 # Entrando com os dados no R
2 Tempo <- c(11, 20, 5, 23, 16, 21, 18, 16, 27, 24, 23, 46, 13, 19, 23, 17, 28,
3           36, 25, 28)
4 Método <- rep(c("A", "B"),each=10)
```

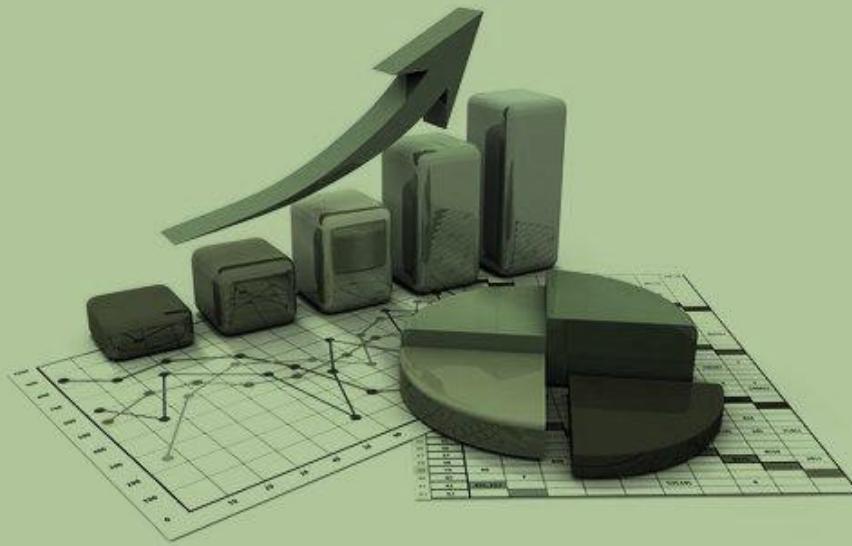
```
5 # Mostrando os dados em forma de data.frame  
6 data.frame(Método,Tempo)  
7  
8 # Gerando o Boxplot comparativo  
9 boxplot(Tempo~Método)
```

Lista 7.14: Comandos do software R

Observação: Na Lista 7.14, dentro do comando `boxplot()`, usou-se "Tempo~Método" para especificar que o Tempo está *em função* do Método.

Tabela 7.12: Tempo (em minutos) para execução do serviço em função do método de treinamento

Método	Tempo
A	11
A	20
A	5
A	23
A	16
A	21
A	18
A	16
A	27
A	24
B	23
B	46
B	13
B	19
B	23
B	17
B	28
B	36
B	25
B	28



8. Medidas de dispersão

As medidas de tendência central resumem a informação contida em um conjunto de dados, mas não contam toda a história. Por causa da variabilidade, a média, a mediana e a moda não são suficientes para descrever um conjunto de dados: informam apenas a tendência central, ou seja, onde está o centro, mas nada dizem sobre a variabilidade.

Para entender esse ponto, imagine dois domicílios:

- No primeiro, moram sete pessoas, todas com 22 anos. A média de idade dos moradores desse domicílio coletivo é, evidentemente, 22 anos:

$$\bar{x} = \frac{22 + 22 + \dots + 22}{7} = 22$$

- No segundo domicílio, também moram sete pessoas: um casal - ela com 17 e ele com 23 anos, dois filhos - um com 2 e outro com 3 anos, a mãe da moça - com 38 anos, um irmão da moça - com 8 anos - e a avó da moça - com 63 anos. A média de idade nesse segundo domicílio também é de 22 anos:

$$\bar{x} = \frac{17 + 23 + 2 + 3 + 38 + 8 + 63}{7} = 22$$

No entanto, a "idade média de 22 anos" descreve bem a situação no primeiro domicílio, mas não no segundo. As medidas de posição são tanto mais descritivas de um conjunto de dados quanto menor é a variabilidade. Então, para representar um conjunto de dados devem ser fornecidas não apenas medidas de posição, mas também uma medida de variabilidade ou dispersão.

8.1 Amplitude

Para medir a variabilidade de um conjunto de dados pode-se fornecer os valores mínimo e máximo do conjunto de dados. Pode-se, também, calcular a amplitude (A) que é definida como a diferença entre o máximo e o mínimo (Equação 8.1):

$$A = X_{\max} - X_{\min} \quad (8.1)$$

■ **Exemplo 8.1** Cinco grupos de alunos submeteram-se a um teste no qual obtiveram as notas apresentadas na Tabela 8.1.

Tabela 8.1: Notas de alunos submetidos a um teste

Grupos	Alunos					Médias
	Aluno 1	Aluno 2	Aluno 3	Aluno 4	Aluno 5	
Grupo A	3	4	5	6	7	5
Grupo B	1	3	5	7	9	5
Grupo C	5	5	5	5	5	5
Grupo D	3	5	5	7	-	5
Grupo E	3	5	5	6	6	5

As amplitudes para cada um dos cinco grupos são dadas a seguir:

- Grupo A: $A = 7 - 3 = 4$;
- Grupo B: $A = 9 - 1 = 8$;
- Grupo C: $A = 5 - 5 = 0$;
- Grupo D: $A = 7 - 3 = 4$;
- Grupo E: $A = 6 - 3 = 3$.

■

A amplitude de variação é uma idéia básica em Estatística, mas um valor discrepante - por ser muito grande ou muito pequeno - aumenta muito a amplitude.

Portanto, o problema em se considerar a amplitude total como medida de dispersão dos dados, é o fato dela levar em consideração em seu cálculo, apenas os valores extremos e não todos os valores. Assim, dois conjuntos de dados podem apresentar a mesma amplitude total, mesmo que tenham dispersão muito diferente. Embora fácil de calcular de interpretar, não deve ser usada normalmente como medida de dispersão.

Na Lista 8.1 são apresentados os comandos do software R para calcular a amplitude.

```

1 # Entrando com os dados no R
2 dados <- c(3, 4, 5, 6, 7)
3
4 # Amplitude
5 A <- diff(range(dados)); A
6
7 # ou, ainda
8 A <- max(dados) - min(dados); A

```

Lista 8.1: Comandos do software R

8.2 Desvio médio absoluto

Outra forma de se medir a variabilidade de uma variável é quantificando a dispersão das observações em relação a um ponto específico na distribuição, em geral, a média. A distância entre os valores

observados e a média denomina-se desvio em relação à média, ou simplesmente desvio, ou seja: $Desvio = x_i - \bar{x}$.

■ **Exemplo 8.2** Considere as notas dos alunos do Grupo A apresentadas na Tabela 8.1 e recorde que a nota média dos alunos deste grupo é $\bar{x} = 5$. Os desvios em relação à média são apresentados na Tabela 8.2.

Tabela 8.2: Desvios em relação à média das notas dos alunos do grupo A

Nota (x_i)	Desvio ($x_i - \bar{x}$)
3	$(3 - 5) = -2$
4	$(4 - 5) = -1$
5	$(5 - 5) = 0$
6	$(6 - 5) = 1$
7	$(7 - 5) = 2$
Total	$\sum(x_i - \bar{x}) = 0$

Observe que a soma dos desvios em relação à média é igual a zero, isto é, $\sum(x_i - \bar{x}) = 0$. Isto sempre ocorrerá para qualquer conjunto de dados. Assim, este somatório não traz nenhuma informação a respeito da variabilidade dos dados. Portanto, para representar a variabilidade dos dados pode-se utilizar a soma dos valores absolutos (módulos) dos desvios, $\sum|x_i - \bar{x}|$, que será sempre maior ou igual a zero. Note que $\sum|x_i - \bar{x}| = 0$ apenas quando não houver variabilidade, ou seja, apenas quando os dados forem todos iguais.

A soma dos módulos dos desvios será tanto maior quanto maior for o número de observações (n). Assim, para obter uma medida de dispersão que não seja afetada pelo número de observações basta dividir a soma dos módulos dos desvios por n . Dessa forma obtém-se o **desvio médio absoluto** (Equação 8.2).

$$d_m = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} \quad (8.2)$$

■ **Exemplo 8.3** Considere as notas dos alunos do Grupo A apresentadas na Tabela 8.1 e recorde que a nota média dos alunos deste grupo é $\bar{x} = 5$. Os módulos dos desvios são apresentados na Tabela 8.3.

Portanto, o desvio médio absoluto para as notas desses alunos é dado por:

$$d_m = \frac{\sum|x_i - \bar{x}|}{n} = \frac{|3 - 5| + \dots + |7 - 5|}{5} = \frac{6}{5} = 1,2 \text{ pontos.}$$

Tabela 8.3: Desvios absolutos das notas dos alunos do grupo A

Nota (x_i)	Desvio ($x_i - \bar{x}$)	Desvio absoluto $ x_i - \bar{x} $
3	$(3 - 5) = -2$	$ 3 - 5 = -2 = 2$
4	$(4 - 5) = -1$	$ 4 - 5 = -1 = 1$
5	$(5 - 5) = 0$	$ 5 - 5 = 0 = 0$
6	$(6 - 5) = 1$	$ 6 - 5 = 1 = 1$
7	$(7 - 5) = 2$	$ 7 - 5 = 2 = 2$
Total	$\sum(x_i - \bar{x}) = 0$	$\sum x_i - \bar{x} = 6$

Na Lista 8.2 são apresentados os comandos do software R para calcular o desvio médio absoluto.

```

1 # Entrando com os dados no R
2 dados <- c(3, 4, 5, 6, 7)
3
4 # Criando uma função para calcular o desvio médio absoluto
5 dm <- function(x) {
6   d <- sum(abs(x-mean(x))) / length(x)
7   return(d)
8 }
9
10 # Desvio médio absoluto
11 dma <- dm(dados); dma

```

Lista 8.2: Comandos do software R

8.3 Variância

8.3.1 Variância para dados brutos

Outra forma de evitar que a soma dos desvios se anule é elevando cada desvio ao quadrado, ou seja, fazendo $\sum(x_i - \bar{x})^2$.

■ **Exemplo 8.4** Considere as notas dos alunos do Grupo A apresentadas na Tabela 8.1 e recorde que a nota média dos alunos deste grupo é $\bar{x} = 5$. Os quadrados dos desvios são apresentados na Tabela 8.4.

Tabela 8.4: Quadrados dos desvios das notas dos alunos do grupo A

Nota (x_i)	Desvio ($x_i - \bar{x}$)	Desvio absoluto $(x_i - \bar{x})^2$
3	$(3 - 5) = -2$	$(3 - 5)^2 = (-2)^2 = 4$
4	$(4 - 5) = -1$	$(4 - 5)^2 = (-1)^2 = 1$
5	$(5 - 5) = 0$	$(5 - 5)^2 = (0)^2 = 0$
6	$(6 - 5) = 1$	$(6 - 5)^2 = (1)^2 = 1$
7	$(7 - 5) = 2$	$(7 - 5)^2 = (2)^2 = 4$
Total	$\sum(x_i - \bar{x}) = 0$	$\sum(x_i - \bar{x})^2 = 10$

■

A partir dos quadrados dos desvios é obtida a variância, que é uma das medidas de dispersão mais utilizadas. Se os dados são provenientes de uma amostra (subconjunto de uma população), a

variância será denotada por s^2 e calculada pela Equação 8.3.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (8.3)$$

■ **Exemplo 8.5** Considere as notas dos alunos do Grupo A e os quadrados dos desvios apresentados na Tabela 8.4 a variância é dada por:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(3-5)^2 + \dots + (7-5)^2}{5-1} = \frac{10}{4} = 2,5 \text{ pontos}^2.$$

■

Observação

Como a variância é calculada usando-se os quadrados dos desvios então a unidade de medida da variância será igual a unidade de medida dos dados elevada ao quadrado. Para o Exemplo 8.5 os dados eram compostos pelas pontuações (*pontos*) de alunos em um teste, então, a unidade de medida da variância será *pontos*².

Na Lista 8.3 são apresentados os comandos do software R para calcular a variância para dados brutos.

```

1 # Entrando com os dados no R
2 dados <- c(3, 4, 5, 6, 7)
3
4 # Variância
5 s2 <- var(dados); s2

```

Lista 8.3: Comandos do software R

Fórmula alternativa da variância

Outra forma de se calcular a variância é usando a Equação 8.4.

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right] \quad (8.4)$$

A fórmula da variância apresentada na Equação 8.4 é resultado de manipulações algébricas da Equação 8.3. Assim, independente de qual fórmula for utilizada o resultado da variância será o mesmo.

■ **Exemplo 8.6** Considere as notas dos alunos do Grupo A apresentadas na Tabela 8.1. Na Tabela 8.5 são apresentados cálculos auxiliares para obter a variância usando a Equação 8.4.

Tabela 8.5: Cálculos auxiliares para obtenção da variância

Aluno	Nota (x_i)	x_i^2
1	3	$3^2 = 9$
2	4	$4^2 = 16$
3	5	$5^2 = 25$
4	6	$6^2 = 36$
5	7	$7^2 = 49$
Total	$\sum x_i = 25$	$\sum x_i^2 = 135$

Assim, obtidos $\sum x_i = 25$ e $\sum x_i^2 = 135$ basta substituí-los na fórmula da variância:

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{5-1} \left[135 - \frac{25^2}{5} \right] = 2,5 \text{ pontos}^2.$$

■

8.3.2 Variância para dados agrupados

Quando os dados estão dispostos em uma tabela de frequências, para se calcular a variância deve-se levar em consideração as frequências (f_i). Assim, a variância para dados agrupados pode ser calculada usando a Equação 8.5:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 f_i}{n-1} \quad (8.5)$$

ou ainda, pode ser calculada usando a Equação 8.6:

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 f_i - \frac{\left(\sum_{i=1}^n x_i f_i \right)^2}{n} \right] \quad (8.6)$$

em que x_i é o valor da variável (para tabela de dados discretos) ou o ponto médio da classe i (para tabela de dados contínuos).

O exemplo a seguir ilustra o cálculo da variância para dados quantitativos discretos agrupados em uma tabela de distribuição de frequências.

■ **Exemplo 8.7** Considere a tabela de distribuição de frequências, com cálculos auxiliares, do número de filhos em idade escolar de vinte funcionários de uma empresa apresentada na Tabela 8.6.

A variância é calculada por:

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} \right] = \frac{1}{20-1} \left[58 - \frac{24^2}{20} \right] = 1,5 \text{ filhos}^2.$$

■

Tabela 8.6: Distribuição de frequências do número de filhos em idade escolar

Número de filhos (x_i)	Frequência (f_i)	$x_i f_i$	$x_i^2 f_i$
0	6	$0 \times 6 = 0$	$0^2 \times 6 = 0$
1	8	$1 \times 8 = 8$	$1^2 \times 8 = 8$
2	4	$2 \times 4 = 8$	$2^2 \times 4 = 16$
3	1	$3 \times 1 = 3$	$3^2 \times 1 = 9$
4	0	$4 \times 0 = 0$	$4^2 \times 0 = 0$
5	1	$5 \times 1 = 5$	$5^2 \times 1 = 25$
Total	$n = \sum f_i = 20$	$\sum x_i f_i = 24$	$\sum x_i^2 f_i = 58$

O exemplo a seguir ilustra o cálculo da variância para dados quantitativos contínuos agrupados em uma tabela de distribuição de frequências.

■ **Exemplo 8.8** Considere a distribuição de frequências, com cálculos auxiliares, dos pesos de 86 indivíduos, apresentada na Tabela 8.7.

Tabela 8.7: Distribuição de frequências dos pesos (em kg) de 86 indivíduos

Peso (Kg)	Frequência (f_i)	Ponto médio (x_i)	$(x_i f_i)$	$(x_i^2 f_i)$
30 ⊂ 40	8	35	$35 \times 8 = 280$	$35^2 \times 8 = 9.800$
40 ⊂ 50	12	45	$45 \times 12 = 540$	$45^2 \times 12 = 24.300$
50 ⊂ 60	15	55	$55 \times 15 = 825$	$55^2 \times 15 = 45.375$
60 ⊂ 70	17	65	$65 \times 17 = 1.105$	$65^2 \times 17 = 71.825$
70 ⊂ 80	14	75	$75 \times 14 = 1.050$	$75^2 \times 14 = 78.750$
80 ⊂ 90	11	85	$85 \times 11 = 935$	$85^2 \times 11 = 79.475$
90 ⊂ 100	9	95	$95 \times 9 = 855$	$95^2 \times 9 = 81.225$
Total	$\sum f_i = 86$	—	$\sum x_i f_i = 5.590$	$\sum x_i^2 f_i = 390.750$

Substituindo os somatórios obtidos na Tabela 8.7 na Equação 8.6 obtém-se:

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} \right] = \frac{1}{86-1} \left[390.750 - \frac{5.590^2}{86} \right] = 322,35 \text{ kg}^2.$$

■

Na Lista 8.5 são apresentados os comandos do software R para calcular a variância para os dados agrupados do Exemplo 8.8.

```

1 # Entrando com os dados no R
2 xi <- c(35, 45, 55, 65, 75, 85, 95) #Pontos médios das classes
3 fi <- c(8, 12, 15, 17, 14, 11, 9) #Frequências
4 dados <- rep(xi,fi)
5
6 # Variância
7 s2 <- var(dados); s2

```

Lista 8.4: Comandos do software R

8.4 Desvio padrão

No cálculo da variância, devido ao fato de se elevar os desvios ao quadrado, a unidade de medida da variância também fica elevada ao quadrado, gerando escalas sem sentido prático. Assim, se a unidade de medida dos dados for metros (m), a unidade de medida da variância será m^2 , se a unidade de medida dos dados for kg , a unidade de medida da variância será kg^2 , etc.

Uma forma de se obter uma medida de dispersão com a mesma unidade de medida dos dados observados é, simplesmente, extrair a raiz quadrada da variância. Fazendo isso obtém-se o desvio padrão, que é a medida de dispersão mais comumente utilizada. O desvio padrão será denotado por s e será dado pela Equação 8.7.

$$s = \sqrt{s^2} \quad (8.7)$$

■ **Exemplo 8.9** Para os dados dos pesos de 86 indivíduos apresentados na Tabela 8.7 obteve-se $s^2 = 322,35 \text{ kg}^2$. Portanto, o desvio padrão dos pesos destes indivíduos será:

$$s = \sqrt{s^2} = \sqrt{322,35} = 17,95 \text{ kg.}$$

Observe que a unidade de medida do desvio padrão é a mesma dos dados, ou seja, kg. ■

Na Lista 8.5 são apresentados os comandos do software R para calcular o desvio padrão para os dados do Exemplo 8.8.

```

1 # Entrando com os dados no R
2 xi <- c(35, 45, 55, 65, 75, 85, 95) #Pontos médios das classes
3 fi <- c(8, 12, 15, 17, 14, 11, 9) #Frequências
4 dados <- rep(xi,fi)
5
6 # Variância
7 s2 <- var(dados); s2
8
9 # Desvio Padrão
10 s <- sqrt(s2); s
11
12 # ou, ainda (usando o comando sd)
13 s <- sd(dados); s

```

Lista 8.5: Comandos do software R

8.5 Coeficiente de variação

A interpretação do desvio padrão depende da ordem de grandeza da variável em estudo. Assim, um desvio padrão igual a 10 pode ser insignificante se os valores típicos observados (x_i) forem em torno de 10.000, mas pode ser muito significativo para um conjunto de dados cujos valores típicos observados sejam em torno de 100.

Logo, pode ser conveniente expressar a variabilidade dos dados de uma variável de modo independente da unidade de medida utilizada, retirando a influência da ordem de grandeza da variável. Tal medida é denominada coeficiente de variação. O coeficiente de variação de Pearson é a razão entre o desvio padrão e a média. Em geral, o resultado é multiplicado por 100, para que o coeficiente de variação seja dado em porcentagem. O coeficiente de variação (cv) é dado pela

Equação 8.8.

$$cv = \left(\frac{s}{\bar{x}} \times 100 \right) \% \quad (8.8)$$

Uma utilidade do coeficiente de variação é fornecer uma medida para o "grau de homogeneidade" de um conjunto de dados. Quanto menor o coeficiente de variação, mais homogêneo é o conjunto de dados (ou seja, mais parecidos os dados são uns com os outros). Em geral, considera-se:

- a) baixa dispersão: $cv < 15\%$;
- b) média dispersão: $15\% \leq cv \leq 30\%$;
- c) alta dispersão: $cv > 30\%$.

O coeficiente de variação também pode ser bastante útil para comparar a variabilidade de duas variáveis ou dois grupos que, a princípio, não são comparáveis por terem, por exemplo, unidades de medidas diferentes.

■ Exemplo 8.10 Na Tabela 8.8 são apresentadas a estatura (cm), o peso (kg) e a idade (anos) de dez indivíduos aleatoriamente selecionados.

Tabela 8.8: Medidas de estatura, peso e idade de dez indivíduos

ID do indivíduo	Estatura (cm)	Peso (kg)	Idade (anos)
1	177	68.0	18.0
2	162	83.0	20.1
3	188	72.0	20.5
4	157	99.9	17.7
5	166	51.0	19.2
6	153	52.0	18.9
7	158	52.0	26.9
8	176	66.5	20.1
9	168	80.0	20.7
10	163	48.0	19.3

Pede-se:

- a) Calcular a média (\bar{x}), a variância (s^2), o desvio padrão (s) e o coeficiente de variação (cv) para as variáveis estatura, peso e idade;
- b) Qual variável apresenta maior variabilidade? Justifique sua resposta;
- c) Classifique a dispersão de cada variável como baixa, média ou alta.

Solução

Solução do item (a)

Para auxiliar nos cálculos das medidas considere a tabela de cálculos auxiliares (Tabela 8.9).

Tabela 8.9: Tabela de cálculos auxiliares

ID do indivíduo	Estatura (x_i)	Peso (y_i)	Idade (z_i)	(x_i^2)	(y_i^2)	(z_i^2)
1	177,00	68,00	18,00	31.329,00	4.624,00	324,00
2	162,00	83,00	20,10	26.244,00	6.889,00	404,01
3	188,00	72,00	20,50	35.344,00	5.184,00	420,25
4	157,00	99,90	17,70	24.649,00	9.980,01	313,29
5	166,00	51,00	19,20	27.556,00	2.601,00	368,64
6	153,00	52,00	18,90	23.409,00	2.704,00	357,21
7	158,00	52,00	26,90	24.964,00	2.704,00	723,61
8	176,00	66,50	20,10	30.976,00	4.422,25	404,01
9	168,00	80,00	20,70	28.224,00	6.400,00	428,49
10	163,00	48,00	19,30	26.569,00	2.304,00	372,49
Total	1.668,00 $\sum x_i$	672,40 $\sum y_i$	201,40 $\sum z_i$	279.264,00 $\sum x_i^2$	47.812,26 $\sum y_i^2$	4.116,00 $\sum z_i^2$

Assim, para a variável estatura (x_i), tem-se:

- $\bar{x} = \frac{\sum x_i}{n} = \frac{1.668,00}{10} = 166,80 \text{ cm};$
- $s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{10-1} \left[279.264,00 - \frac{1.668,00^2}{10} \right] = 115,73 \text{ cm}^2;$
- $s = \sqrt{s^2} = \sqrt{115,73} = 10,76 \text{ cm};$
- $cv = \left(\frac{s}{\bar{x}} \times 100 \right) \% = \left(\frac{10,76}{166,80} \times 100 \right) \% = 6,45\%.$

Fazendo cálculos análogos para as variáveis "peso" e "idade" são obtidos a média, a variância, o desvio padrão e o coeficiente de variação de todas as variáveis, que são apresentados na Tabela 8.10.

Tabela 8.10: Medidas descritivas das variáveis estatura, peso e idade

Medida	Variável		
	Estatura	Peso	Idade
\bar{x}	166,80	67,24	20,14
s^2	115,73	288,90	6,64
s	10,76	17,00	2,58
cv	6,45%	25,28%	12,81%

Solução do item (b)

A variável que apresentou maior variabilidade foi a variável peso pois ela apresentou o maior coeficiente de variação: $cv = 25,28\%$.

Observação: Apesar de a variância e o desvio padrão também terem sido maiores para a variável

peso a conclusão de que esta variável tem maior variabilidade do que as outras não pode ser tomada utilizando-se a variância ou do desvio padrão, pois, estas medidas não servem para comparar variáveis com diferentes unidades de medida ou com grandezas diferentes. Para essa finalidade deve-se, necessariamente, utilizar o coeficiente de variação.

Solução do item (c)

- **Estatura:** dispersão baixa $cv < 15\%$;
- **Peso:** dispersão média $15\% \leq cv \leq 30\%$;
- **Idade:** dispersão baixa $cv < 15\%$.

■

Na Lista 8.6 são apresentados os comandos do software R para calcular os coeficientes de variação do Exemplo 8.10.

```

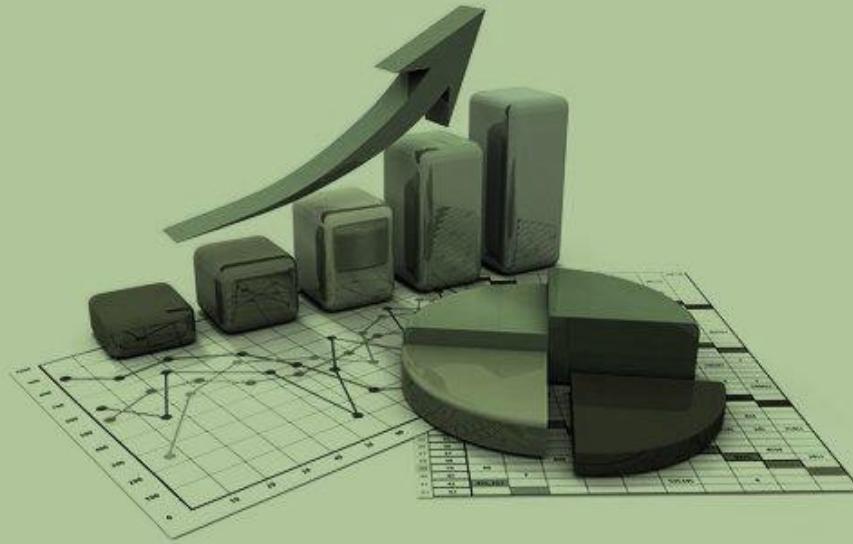
1 # Entrando com os dados no R
2 Estatura <- c(177, 162, 188, 157, 166, 153, 158, 176, 168, 163)
3 Peso <- c(68.0, 83.0, 72.0, 99.9, 51.0, 52.0, 52.0, 66.5, 80.0, 48.0)
4 Idade <- c(18.0, 20.1, 20.5, 17.7, 19.2, 18.9, 26.9, 20.1, 20.7, 19.3)
5
6 # Coeficiente de variação para a variável Estatura
7 cv1 <- 100*sd(Estatura) /mean(Estatura)
8 cv1
9
10 # Coeficiente de variação para a variável Peso
11 cv2 <- 100*sd(Peso) /mean(Peso)
12 cv2
13
14 # Coeficiente de variação para a variável Idade
15 cv3 <- 100*sd(Idade) /mean(Idade)
16 cv3

```

Lista 8.6: Comandos do software R

8.6 Propriedades das Medidas de Dispersão

1. todas as medidas de dispersão são não negativas;
2. somando-se ou subtraindo-se uma mesma constante não nula (k) a todas as observações, as medidas de dispersão não se alteram, pois ocorre apenas uma translação dos valores, ou seja, $var(X \pm k) = var(X)$;
3. quando multiplica-se ou divide-se todos os valores de uma variável (X) por uma constante (k), a sua **variância** ($var(X)$) fica multiplicada ou dividida pelo **quadrado da constante**, ou seja, $var(k \times X) = k^2 \times var(X)$ e $var\left(\frac{X}{k}\right) = \frac{var(X)}{k^2}$;
4. quando multiplica-se ou divide-se todos os valores de uma variável (X) por uma constante (k), o seu **desvio padrão** ($dp(X)$) fica multiplicado ou dividido pela constante, ou seja, $dp(k \times X) = k \times dp(X)$ e $dp\left(\frac{X}{k}\right) = \frac{dp(X)}{k}$.



9. Assimetria e curtose

9.1 Assimetria

Numa distribuição estatística, a assimetria é o quanto sua curva de frequências se desvia ou se afasta da posição simétrica. Pode-se analisar a assimetria de uma distribuição de acordo com as relações entre suas medidas de moda, média e mediana.

Graficamente, tem-se um eixo de referência ou eixo de simetria, que é traçado sobre o valor da média da distribuição. Na Figura 9.1 podemos observar uma distribuição simétrica com o eixo de simetria no centro da distribuição.

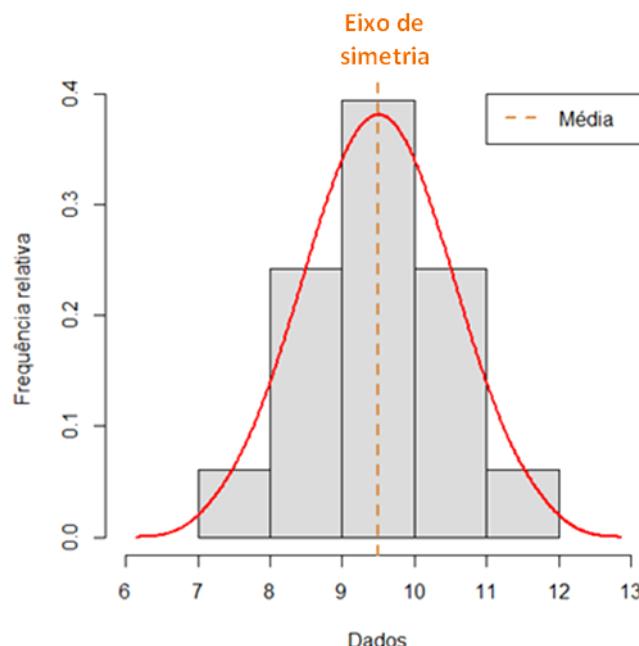


Figura 9.1: Distribuição simétrica com eixo de simetria no centro

Sempre que a curva da distribuição se afastar do referido eixo, será considerada como tendo um certo grau de afastamento, que é considerado como uma assimetria da distribuição. Ou seja, assimetria é o grau de afastamento que uma distribuição apresenta do seu eixo de simetria.

Pode-se caracterizar uma distribuição de frequência em:

- Distribuição simétrica (ou assimetria nula);
- Distribuição assimétrica à direita (ou positiva);
- Distribuição assimétrica à esquerda (ou negativa).

Distribuição simétrica (ou assimetria nula)

Uma distribuição é dita simétrica (Figura 9.2) quando apresenta o mesmo valor para a moda, a média e a mediana, ou seja: $\bar{x} = m_d = m_o$.

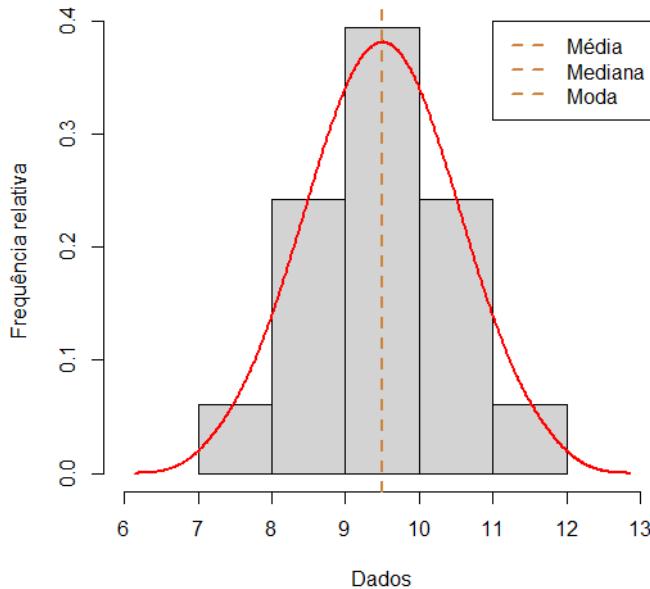


Figura 9.2: Distribuição simétrica

Distribuição assimétrica à direita (ou positiva)

Quando a cauda da curva da distribuição declina para direita, tem-se uma distribuição com curva assimétrica positiva (Figura 9.3). Neste caso, tem-se a relação: $\bar{x} > m_d > m_o$.

Distribuição assimétrica à esquerda (ou negativa)

Analogamente, quando a cauda da curva da distribuição declina para esquerda, tem-se uma distribuição com curva assimétrica negativa (Figura 9.4). Neste caso, tem-se a relação: $\bar{x} < m_d < m_o$.

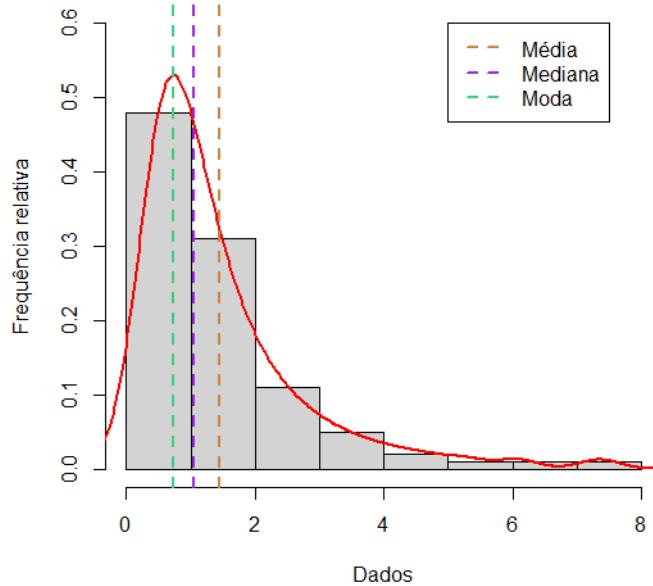


Figura 9.3: Distribuição assimétrica à direita

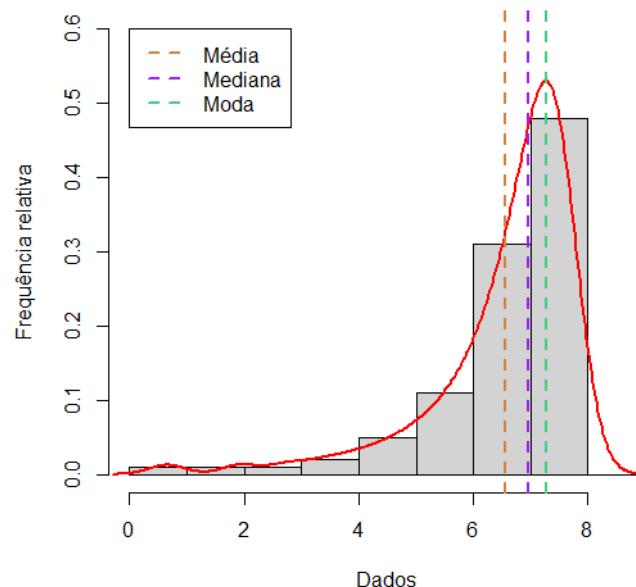


Figura 9.4: Distribuição assimétrica à esquerda

Coeficiente momento de assimetria

Outra forma de verificar a assimetria de uma distribuição é por meio de um coeficiente de assimetria. Existem vários coeficientes de assimetria e, dentre eles, pode-se citar o coeficiente momento de assimetria, calculado com base nos momentos centrais de segunda e terceira ordem, definido pela Equação 9.1.

$$AS = \frac{m_3}{(\sqrt{m_2})^3} \quad (9.1)$$

em que:

$$m_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad \text{e} \quad m_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n},$$

para dados brutos, ou, ainda:

$$m_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 f_i}{n} \quad \text{e} \quad m_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3 f_i}{n},$$

se os dados estiverem agrupados em uma distribuição de frequências.

A interpretação do coeficiente de assimetria é:

- $AS = 0$: a distribuição é simétrica;
- $AS > 0$: a distribuição é assimétrica positiva (à direita);
- $AS < 0$: a distribuição é assimétrica negativa (à esquerda).

■ **Exemplo 9.1** Considere os dados brutos dos salários (em x sal. mín.) de 36 indivíduos:

5,73	13,60	13,23	8,46	17,26	16,22	8,74	23,30	7,39
11,06	13,85	8,12	15,99	10,76	6,26	9,80	5,25	9,77
19,40	10,53	11,59	14,69	8,95	9,35	4,56	4,00	9,13
14,71	12,00	7,59	7,44	6,66	12,79	18,75	6,86	16,61

A média e a mediana dos dados brutos dos salários são, respectivamente, $\bar{x} = 11,12$ e $m_d = 10,17$. Assim, tem-se a relação $\bar{x} > m_d$ e, portanto, a distribuição dos salários é assimétrica à direita. **Observação:** A moda não foi apresentada neste exemplo pois para dados contínuos deve-se agrupar os dados em uma distribuição de frequências para estimar a moda.

Para calcular o coeficiente momento de assimetria deve-se, primeiro, calcular os momentos centrais de segunda e terceira ordem:

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{736,57}{36} = 20,46 \quad \text{e} \quad m_3 = \frac{\sum (x_i - \bar{x})^3}{n} = \frac{2.084,59}{36} = 57,91.$$

Portanto, o coeficiente momento de assimetria é obtido por:

$$AS = \frac{m_3}{(\sqrt{m_2})^3} = \frac{57,91}{(\sqrt{20,46})^3} = 0,63.$$

Como $AS > 0$ então a distribuição dos salários é assimétrica à direita.

Pode-se também observar a assimetria desta distribuição por meio de gráficos. Na Figura 9.5 são apresentados um histograma e um boxplot dos salários nos quais fica evidente que a distribuição é assimétrica à direita.

■

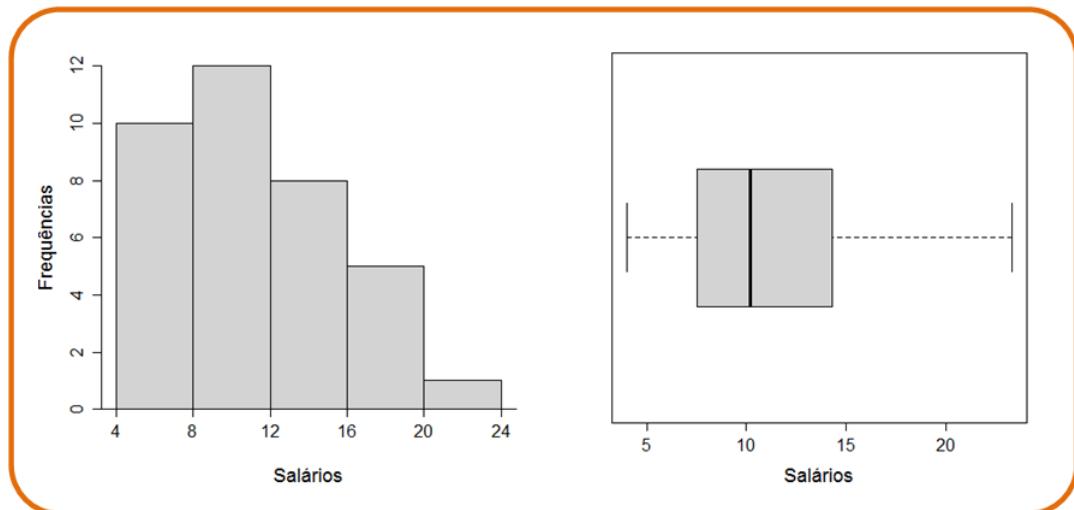


Figura 9.5: Histograma e boxplot dos salários

Na Lista 9.1 são apresentados os comandos do software R para calcular o coeficiente momento de assimetria para os dados de salários do Exemplo 9.1.

```

1 # Entrando com os dados no R
2 dados <- c(5.73, 13.60, 13.23, 8.46, 17.26, 16.22, 8.74, 23.30, 7.39,
3   11.06, 13.85, 8.12, 15.99, 10.76, 6.26, 9.80, 5.25, 9.77,
4   19.40, 10.53, 11.59, 14.69, 8.95, 9.35, 4.56, 4.00, 9.13,
5   14.71, 12.00, 7.59, 7.44, 6.66, 12.79, 18.75, 6.86, 16.61)
6
7 # Carregando o pacote "moments" (precisa instalar)
8 library(moments)
9
10 # Coeficiente momento de assimetria
11 skewness(dados)
```

Lista 9.1: Comandos do software R

9.2 Curtose

A curtose é uma medida do grau de achatamento da distribuição quando comparada ao de uma distribuição conhecida como distribuição normal (que será vista na disciplina de Probabilidade). Na Figura 9.6 pode-se obserar distribuições com diferentes graus de curtose.

Existem coeficientes que podem ser utilizados para determinar o grau de curtose de uma distribuição de frequências. Dentre eles pode-se citar o coeficiente momento de curtose. Este coeficiente é definido na Equação 9.2, e é dado pelo quociente entre o momento centrado de quarta ordem e o quadrado do momento centrado de segunda ordem.

$$k_m = \frac{m_4}{(m_2)^2} \quad (9.2)$$

em que:

$$m_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad \text{e} \quad m_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n},$$

para dados brutos, ou, ainda:

$$m_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 f_i}{n} \quad \text{e} \quad m_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 f_i}{n},$$

se os dados estiverem agrupados em uma distribuição de frequências.

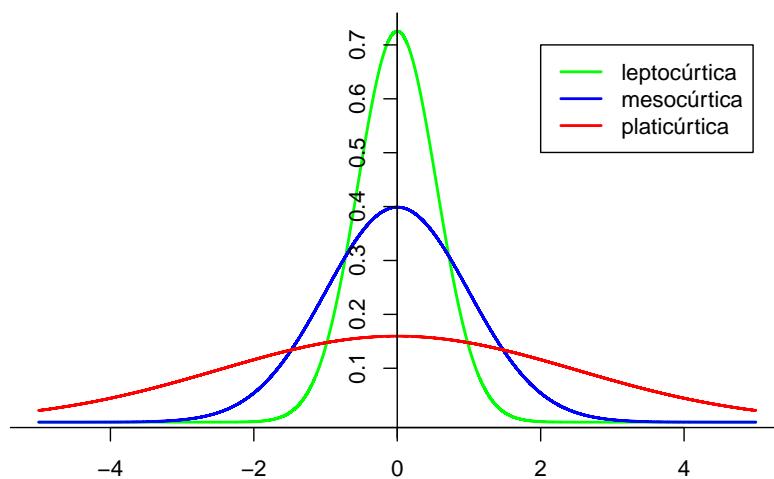


Figura 9.6: Distribuições com diferentes graus de curtose

Para a distribuição Normal o valor do coeficiente de curtose é 3, ou seja, é uma distribuição mesocúrtica. Coeficientes maiores que 3, representam as distribuições mais "afiladas" que a distribuição Normal, ou seja, leptocúrticas. E distribuições com coeficientes de curtose menores do que 3 representam as distribuições mais achatadas do que a Normal, ou seja, platicúrticas. Resumidamente tem-se a seguinte interpretação do coeficiente momento de curtose:

- Se $k_m < 3$, a curva ou distribuição é platicúrtica;
- Se $k_m = 3$, a curva ou distribuição é mesocúrtica;
- Se $k_m > 3$, a curva ou distribuição é leptocúrtica.

■ **Exemplo 9.2** Considere os dados brutos das idades (em anos) de 36 indivíduos:

21,52	23,11	40,52	31,70	36,14	26,98	31,90	30,88	35,07
17,58	27,82	31,54	18,99	24,77	19,70	29,70	29,41	31,48
25,66	25,34	20,96	29,63	30,46	32,31	35,63	31,69	32,19
24,90	29,35	39,00	29,26	27,10	22,61	31,16	25,32	20,35

Para calcular o coeficiente momento de curtose deve-se, primeiro, calcular os momentos centrais de segunda e quarta ordem:

$$m_2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{1.088,89}{36} = 30,25 \quad \text{e} \quad m_4 = \frac{\sum(x_i - \bar{x})^4}{n} = \frac{82.885,22}{36} = 2.302,37.$$

Portanto, o coeficiente momento de curtose é obtido por:

$$k_m = \frac{m_4}{(m_2)^2} = \frac{2.302,37}{(30,25)^2} = 2,52.$$

Como $k_m < 3$ então a distribuição dos salários é platicúrtica.

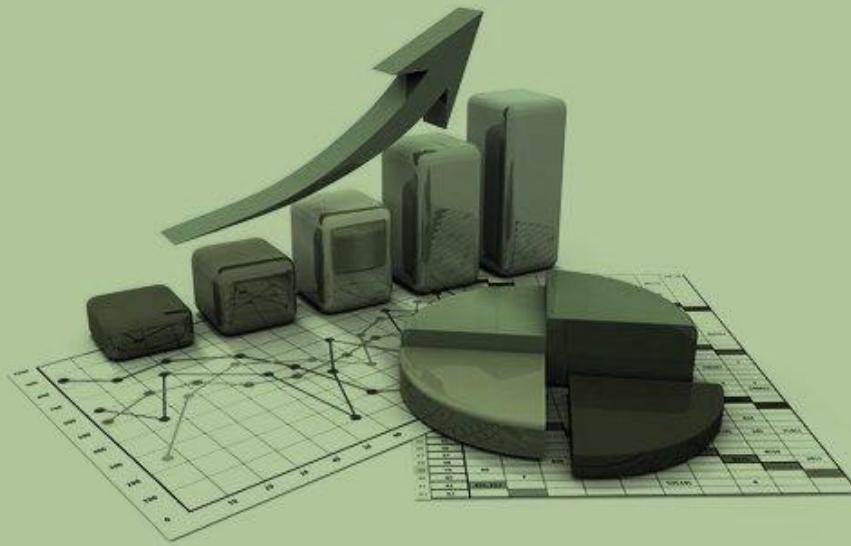
■

Na Lista 9.2 são apresentados os comandos do software R para calcular o coeficiente momento de curtose para os dados de idades do Exemplo 9.2.

```

1 # Entrando com os dados no R
2 dados <- c(21.52, 23.11, 40.52, 31.70, 36.14, 26.98, 31.90, 30.88, 35.07,
3   17.58, 27.82, 31.54, 18.99, 24.77, 19.70, 29.70, 29.41, 31.48,
4   25.66, 25.34, 20.96, 29.63, 30.46, 32.31, 35.63, 31.69, 32.19,
5   24.90, 29.35, 39.00, 29.26, 27.10, 22.61, 31.16, 25.32, 20.35)
6
7
8 # Carregando o pacote "moments" (precisa instalar)
9 library(moments)
10
11 # Coeficiente momento de curtose
12 kurtosis(dados)
```

Lista 9.2: Comandos do software R



10. Análise Bivariada

Até agora foi visto como organizar e resumir informações pertinentes a uma única variável, mas frequentemente, existe o interesse em analisar o comportamento conjunto de duas ou mais variáveis.

10.1 Variáveis qualitativas

10.1.1 Tabelas de contingência e gráficos da distribuição conjunta

Tabelas de contingência são tabelas de dupla entrada que representam a distribuição conjunta de duas variáveis qualitativas.

■ **Exemplo 10.1** Suponha que se queira analisar o comportamento conjunto das variáveis X : grau de instrução e Y : região de procedência de 36 indivíduos cujos dados são apresentados na Tabela 10.1.

A distribuição de frequências conjunta das variáveis X e Y é obtida contando-se o número de indivíduos que se enquadram simultaneamente em um nível da variável X e em um nível da variável Y . Por exemplo, o número de indivíduos que vêm da capital e tem ensino fundamental é igual a 4 pois quatro indivíduos se enquadram nessas duas categorias simultaneamente. Essas frequências são organizadas em uma tabela de dupla entrada (Tabela 10.2), chamada de tabela de contingência.

Observe na Tabela 10.2 que a linha dos totais (valores: 12, 18 e 6) fornece a distribuição da variável X e a coluna dos totais (valores: 11, 12 e 13) fornece a distribuição da variável Y . As distribuições assim obtidas são chamadas de distribuições marginais de X e Y .

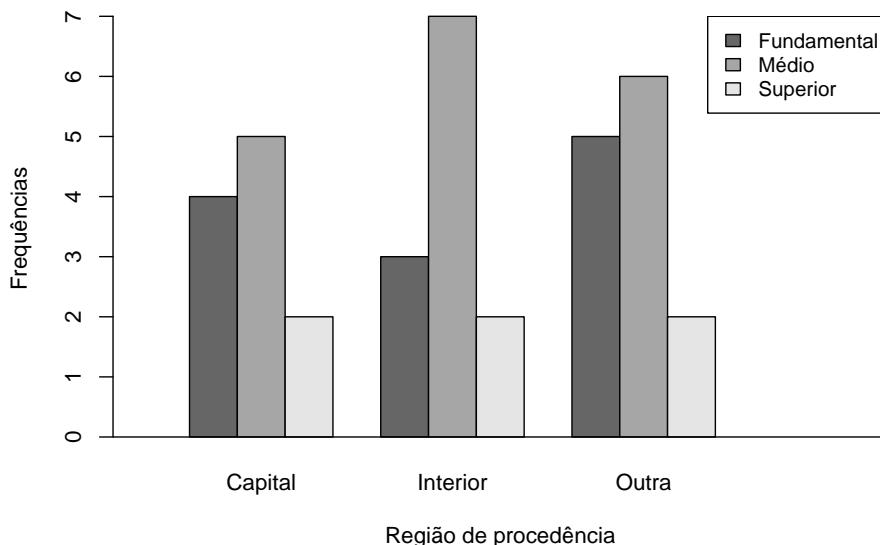
A distribuição conjunta de duas variáveis pode também ser feita graficamente. Na Figura 10.1 é apresentado um gráfico de colunas da distribuição conjunta das variáveis grau de instrução (X) e região de procedência (Y). ■

Tabela 10.1: Região de procedência e grau de instrução de 36 indivíduos

ID	Região proced.	Grau instrução	ID	Região proced.	Grau instrução	ID	Região proced.	Grau instrução
1	Capital	Fundamental	13	Interior	Médio	25	Outra	Fundamental
2	Interior	Fundamental	14	Capital	Médio	26	Outra	Médio
3	Interior	Médio	15	Capital	Médio	27	Capital	Fundamental
4	Outra	Fundamental	16	Outra	Médio	28	Outra	Fundamental
5	Interior	Médio	17	Interior	Superior	29	Outra	Fundamental
6	Outra	Médio	18	Interior	Médio	30	Capital	Médio
7	Outra	Fundamental	19	Interior	Fundamental	31	Outra	Médio
8	Capital	Fundamental	20	Outra	Médio	32	Capital	Médio
9	Outra	Fundamental	21	Capital	Superior	33	Interior	Médio
10	Capital	Fundamental	22	Capital	Superior	34	Interior	Médio
11	Interior	Fundamental	23	Outra	Superior	35	Capital	Médio
12	Interior	Superior	24	Interior	Médio	36	Outra	Superior

Tabela 10.2: Distribuição conjunta do grau de instrução (X) e região de procedência (Y)

$Y \backslash X$				Total
	Fundamental	Médio	Superior	
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Figura 10.1: Distribuição conjunta das frequências das variáveis grau de instrução (X) e região de procedência (Y)

Na Lista 10.1 são apresentados os comandos do software R para gerar a tabela e o gráfico da distribuição conjunta do Exemplo 10.1.

```

1 # Entrando com os dados no R
2 dados <- read.table("C:caminho_do_diretorio/data6.txt", h=T)
3 dados
4
5 # Tabela de distribuição conjunta
6 tabela <- table(dados$Região, dados$Instrução); tabela
7
8 # Distribuições marginais (linha e coluna dos totais)
9 addmargins(tabela)
10
11 # Plotando o gráfico da distribuição conjunta
12 barplot(t(tabela), beside=T, col=c("gray40", "gray65", "gray90")),
13   xlab="Região de procedência", ylab="Frequências", xlim=c(0,16))
14 abline(h=0)
15
16 # Adicionando as legendas (nomes das colunas por cores)
17 legend("topright", legend=colnames(tabela), cex=0.9,
18   fill=c("gray40", "gray65", "gray90"))

```

Lista 10.1: Comandos do software R

Observações:

- (a) no comando `barplot()` foi usado o argumento `t(tabela)` para transpor a matriz da distribuição conjunta de forma que as regiões de procedência fiquem nas colunas e apareçam como as categorias do gráfico de colunas;
- (b) o arquivo de dados utilizado na Lista 10.1 pode ser baixado em <https://drive.google.com/drive/folders/1tCzWnInYGOdpC0qQsJ2EvJLDqLQ7z8aB?usp=sharing>.

10.1.2 Associação entre Variáveis Qualitativas

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas, isto é, quando se quer conhecer o grau de dependência entre elas.

- **Exemplo 10.2** Considere a distribuição conjunta do sexo e escolha do curso de 200 alunos de uma faculdade (Tabela 10.3).

Tabela 10.3: Distribuição conjunta de alunos segundo o sexo (X) e a escolha do curso (Y)

\backslash	X	Masculino	Feminino	Total
Y				
Economia		85	35	120
Administração		55	25	80
Total		140	60	200

Inicialmente é muito difícil tirar alguma conclusão sobre a existência de associação entre as duas variáveis devido à diferença entre os totais marginais. Deve-se, então, calcular as proporções ou porcentagens segundo as linhas ou segundo as colunas para que se possa fazer comparações. Calculando as porcentagens em relação aos totais das linhas obtém-se a Tabela 10.4.

Tabela 10.4: Porcentagens em relação aos totais das linhas da distribuição conjunta do sexo (X) e da escolha do curso (Y)

$X \backslash Y$	Masculino	Feminino	Total
Economia	$85/120 = 70,8\%$	$35/120 = 29,2\%$	$120/120 = 100\%$
Administração	$55/80 = 68,8\%$	$25/80 = 31,2\%$	$80/80 = 100\%$
Total	$140/200 = 70,0\%$	$60/200 = 30,0\%$	$200/200 = 100\%$

Pode-se ver na distribuição marginal da variável sexo (linha dos totais), que é independente do curso escolhido, que 70% dos alunos são do sexo masculino e 30% dos alunos são do sexo feminino. Se não houvesse dependência entre as variáveis esperaria-se as mesmas proporções de homens (70%) e de mulheres (30%) para cada um dos cursos escolhidos. No curso de economia as porcentagens são de 70,8% de homens e 29,2% de mulheres e no curso de administração as porcentagens são de 68,8% de homens e 31,2% de mulheres. Como as porcentagens de homens e mulheres nos dois cursos são muito próximas **parece não haver dependência** entre as duas variáveis (sexo e escolha do curso).

Outra forma de avaliar a dependência entre duas variáveis é por meio de gráficos. Na Figura 10.2 pode-se observar a distribuição conjunta das variáveis sexo e escolha do curso em termos das porcentagens de homens e mulheres em cada curso. Observe que como as porcentagens de homens e mulheres no curso de economia são muito parecidas com as porcentagens de homens e mulheres no curso de administração então parece que a escolha do curso independe do sexo, ou seja, **parece não haver dependência** entre as variáveis sexo e escolha do curso.

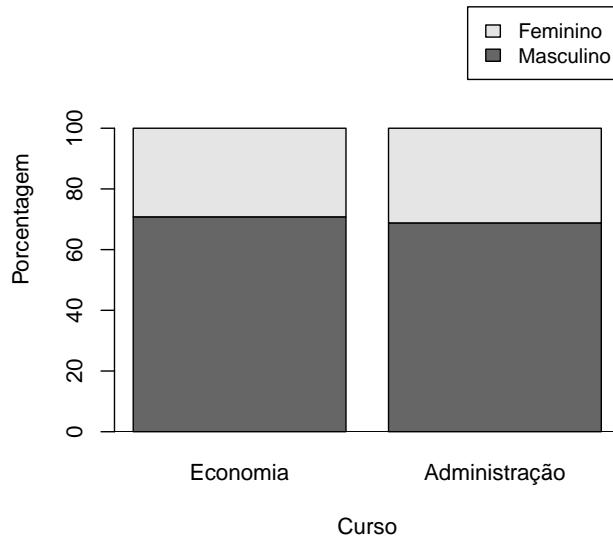


Figura 10.2: Distribuição conjunta do sexo (X) e da escolha do curso (Y)

■ **Exemplo 10.3** Considere, agora, um problema semelhante ao do exemplo anterior, porém envolvendo alunos dos cursos de Física e Ciências Sociais. A distribuição conjunta das variáveis sexo e escolha do curso são apresentadas na Tabela 10.5.

Tabela 10.5: Porcentagens em relação aos totais das linhas da distribuição conjunta do sexo (X) e da escolha do curso (Y)

$Y \backslash X$	Masculino	Feminino	Total
Física	100	20	120
Ciências sociais	40	40	80
Total	140	60	200

Calculando as porcentagens segundo os totais das linhas obtém-se a Tabela 10.6.

Tabela 10.6: Porcentagens, segundo os totais das linhas, da distribuição conjunta do sexo (X) e do curso escolhido (Y).

$Y \backslash X$	Masculino	Feminino	Total
Física	$100/120 = 83,3\%$	$20/120 = 16,7\%$	$120/120 = 100\%$
Ciências sociais	$40/80 = 50,0\%$	$40/80 = 50,0\%$	$80/80 = 100\%$
Total	$140/200 = 70,0\%$	$60/200 = 30,0\%$	$200/200 = 100\%$

Pode-se ver que no curso de Física as porcentagens são de 83,3% de homens e 16,7% de mulheres e no curso de Ciências sociais as porcentagens são de 50,0% de homens e 50,0% de mulheres. Pode-se observar também pela Figura 10.3 que as porcentagens de homens e mulheres no curso de Física são muito diferentes das porcentagens de homens e mulheres no curso de Ciências sociais, ou seja, **parece haver dependência** entre as duas variáveis.

■

Na Lista 10.2 são apresentados os comandos do software R para obter as porcentagens segundo os totais das linhas e o gráfico da distribuição conjunta em termos das porcentagens de homens e mulheres em cada curso.

```

1 # Entrando com a tabela da distribuição conjunta no R
2 tabela <- matrix(c(100,20,40,40), 2, 2, byrow=T)
3 rownames(tabela) <- c("Física", "Ciências sociais")
4 colnames(tabela) <- c("Masculino", "Feminino")
5 tabela
6
7 # Adicionando a linha dos totais
8 tabela2 <- addmargins(tabela,1)
9 tabela2
10
11 # Tabela de proporções (em porcentagens)
12 tabprop <- round(prop.table(tabela2,1)*100, 1)
13 tabprop
14

```

```

15 # Coluna dos totais das porcentagens
16 addmargins(tabprop,2)
17 tabprop <- tabprop[-3,]
18
19 # Plotando o gráfico da distribuição conjunta
20 barplot(t(tabprop), xlab="Curso", ylab="Porcentagem", axes=F,
21         ylim=c(0,140), col=c("gray40","gray90"))
22 axis(2, at=seq(0,100,20)); abline(h=0)
23
24 # Adicionando as legendas (nomes das colunas por cores)
25 legend("topright", legend=rev(colnames(tabela2)), cex=0.9,
26        fill=rev(c("gray40","gray90")))

```

Lista 10.2: Comandos do software R

Observação: Dentro do comando `legend()`, nos argumentos `legend` e `fill`, foi utilizado o comando `rev` para inverter a ordem dos nomes e das cores da legenda, de modo que o sexo feminino fosse apresentado primeiro do que o sexo masculino na legenda.

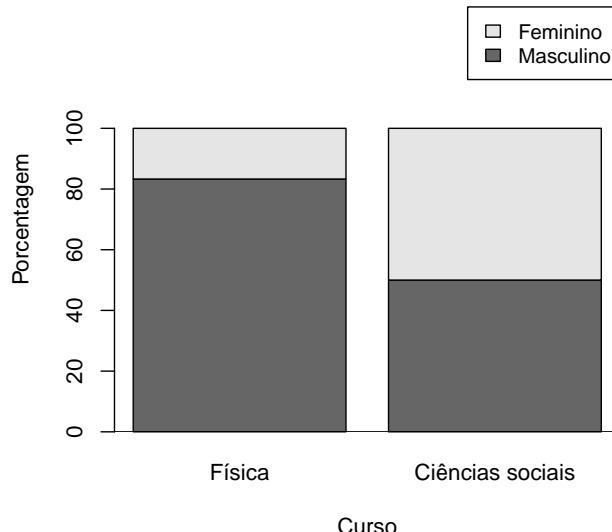


Figura 10.3: Distribuição conjunta do sexo (X) e da escolha do curso (Y)

10.1.3 Qui-quadrado (χ^2)

Define-se o Qui-quadrado de Pearson pela Equação 10.1.

$$\sum_{i=1}^r \sum_{j=1}^s \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (10.1)$$

em que:

- r é o número de linhas da tabela de contingência;
- s é o número de colunas da tabela de contingência;
- o_{ij} são as frequências observadas na distribuição conjunta;

- e_{ij} são as frequências esperadas, ou seja, valores esperados das frequências caso não existisse associação entre as variáveis. O cálculo das frequências esperadas é feito por:

$$e_{ij} = \frac{(\text{Total da linha } i) \times (\text{Total da coluna } j)}{\text{Total de observações}}.$$

O qui-quadrado é uma medida de afastamento global entre as frequências observadas e esperadas e quanto maior seu valor maior é o grau de associação entre as variáveis.

■ **Exemplo 10.4** Considere a distribuição conjunta do sexo e da escolha do curso apresentada no Exemplo 10.3 (Tabela 10.5).

Recordando a Tabela 10.5				
\backslash	X	Masculino	Feminino	Total
Y				
Física		100	20	120
Ciências sociais		40	40	80
Total		140	60	200

As frequências no interior da Tabela 10.5 são as frequências observadas. As frequências esperadas são calculadas por:

$$\begin{array}{ll} e_{11} = \frac{120 \times 140}{200} = 84 & e_{12} = \frac{120 \times 60}{200} = 36 \\ \hline e_{21} = \frac{80 \times 140}{200} = 56 & e_{22} = \frac{80 \times 60}{200} = 24 \end{array}$$

Na Tabela 10.8 são apresentadas as frequências observadas (o_{ij}) e esperadas (e_{ij}) segundo o sexo (X) e a escolha do curso (Y).

Tabela 10.8: Frequências observadas (o_{ij}) e esperadas (e_{ij}) segundo o sexo (X) e a escolha do curso (Y)

\backslash	X	Masculino	Feminino	Total
Y				
Física		100 (84)	20 (36)	120
Ciências sociais		40 (56)	40 (24)	80
Total		140	60	200

Portanto o qui-quadrado é calculado como:

$$\sum_{i=1}^r \sum_{j=1}^s \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{(100 - 84)^2}{84} + \frac{(20 - 36)^2}{36} + \frac{(40 - 56)^2}{56} + \frac{(40 - 24)^2}{24} = 25,4.$$

■

10.1.4 Coeficiente de contingência

Pearson definiu, ainda, uma medida de associação baseada no qui-quadrado, chamada de coeficiente de contingência (de Pearson), dado pela Equação 10.2.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \quad (10.2)$$

Quanto mais próximo de zero o valor de C estiver, menor será a associação entre as variáveis X e Y e quanto mais próximo de um o valor de C estiver, maior será a associação entre as duas variáveis. Contudo, o coeficiente acima nunca atinge o valor 1. O valor máximo de C depende de r e s (número de linhas e colunas, respectivamente). Para evitar esse inconveniente, costuma-se definir um outro coeficiente de contingência (de Tschuprow), que é um coeficiente de contingência corrigido, dado pela Equação 10.3.

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(s-1)}}, \quad (10.3)}$$

que pode atingir o máximo igual a 1 se $r = s$ (número de linhas tabela igual ao número de colunas).

■ **Exemplo 10.5** Considere o Exemplo 10.4. O valor obtido do qui-quadrado foi $\chi^2 = 25,4$. Os coeficientes de contingência são calculados por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{25,4}{25,4 + 200}} = 0,336,$$

e

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(s-1)}}} = \sqrt{\frac{25,4/200}{\sqrt{(2-1)(2-1)}}} = 0,3563.$$

indicando uma associação "moderada" entre as variáveis sexo (X) e escolha do curso (Y). ■

Na Lista 10.3 são apresentados os comandos do software R para obter o coeficiente de contingência.

```

1 # Entrando com a distribuição conjunta no R
2 tabela <- matrix(c(100, 20, 40, 40), 2, 2, byrow=T)
3
4 # Adicionando os nomes das linhas e colunas
5 rownames(tabela) <- c("Física", "Ciências sociais")
6 colnames(tabela) <- c("Masculino", "Feminino")
7 tabela
8

```

```

9 # Qui-quadrado usando o comando "chisq.test()"
10 chisq.test(tabela,correct=FALSE)
11
12 # Carregando o pacote vcd (precisa instalar)
13 library(vcd)
14
15 # Diversos qui-quadrados e coeficientes de contingência
16 assocstats(tabela)
17
18 # Qui-quadrado de Pearson
19 X2 <- assocstats(tabela)[[2]][2,1]; X2
20
21 # Coeficiente de contingência de Pearson (C)
22 C <- assocstats(tabela)[[4]]; C
23
24 # Coeficiente de contingência de Tschuprow (T)
25 n <- sum(tabela)
26 r <- nrow(tabela)
27 s <- ncol(tabela)
28 T <- sqrt((X2/n)/sqrt((r-1)*(s-1))); T

```

Lista 10.3: Comandos do software R

Observação: No comando `chisq.test()` usou-se o argumento `correct=FALSE` para que não fosse utilizada a correção de Yates para o teste Qui-quadrado em tabelas 2×2 . Como, nesta disciplina, ainda não estão sendo considerados aspectos referentes à distribuição de probabilidades de estatísticas, que é o motivo pelo qual a correção de Yates é aplicada, optou-se por não utilizar esta correção no cálculo do χ^2 .

10.2 Variáveis quantitativas

10.2.1 Gráfico de dispersão

Um recurso bastante útil para se verificar a associação entre duas variáveis quantitativas é o gráfico de dispersão, que será introduzido por meio de exemplos.

■ **Exemplo 10.6** Na Tabela 10.9 são apresentados o tempo de serviço (X), em anos, e o número de clientes (Y) de agentes de uma companhia de seguros.

Na Figura 10.4 é apresentado o gráfico de dispersão das variáveis X e Y . Nesse tipo de gráfico são apresentados os pares (x, y) em um plano cartesiano.

Pode-se observar por este gráfico que parece haver uma associação entre as duas variáveis porque à medida que aumenta o tempo de serviço o número de clientes também aumenta, e essa relação parece ocorrer de forma linear.

Tabela 10.9: Anos de serviço (X) e o número de clientes (Y) de agentes de uma companhia de seguros

ID do agente	Anos de serviço (X)	Número de clientes (Y)
1	2	48
2	3	50
3	4	56
4	5	52
5	4	43
6	6	60
7	7	62
8	8	58
9	8	64
10	10	72

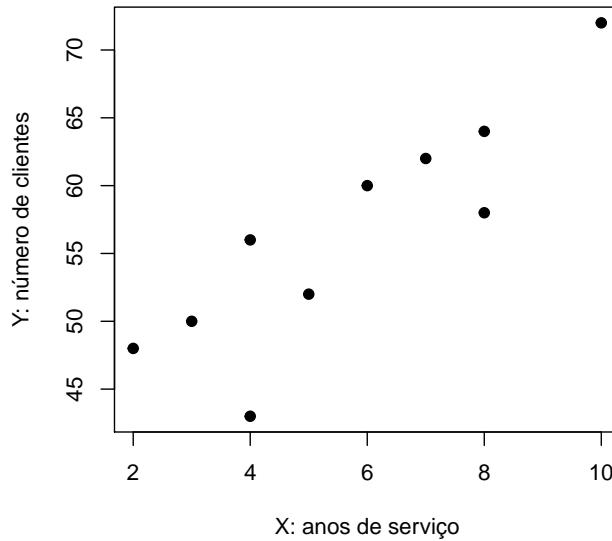


Figura 10.4: Gráfico de dispersão das variáveis X : anos de serviço e Y : número de clientes

Na Lista 10.5 são apresentados os comandos do software R para obter o gráfico de dispersão X versus Y .

```

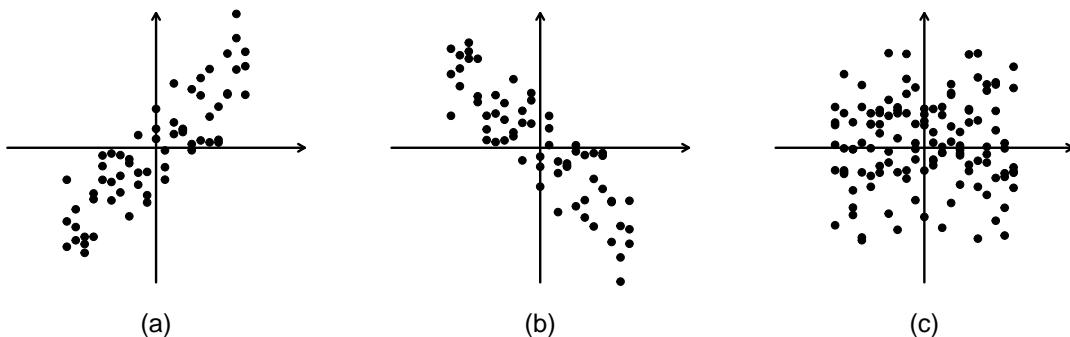
1 # Entrando com os dados no R
2 X <- c(2, 3, 4, 5, 4, 6, 7, 8, 8, 10)
3 Y <- c(48, 50, 56, 52, 43, 60, 62, 58, 64, 72)
4
5 # Gráfico de dispersão X versus Y
6 plot(X, Y, pch=19, xlab="X: tempo de serviço, em anos",
7      ylab="Y: número de clientes")

```

Lista 10.4: Comandos do software R

Na Figura 10.5 pode-se observar possíveis tipos de associação (correlação) linear entre as variáveis. Na Figura 10.5(a) existe uma associação linear positiva, ou seja, ao passo que uma variável aumenta a outra também aumenta. Na Figura 10.5(b) existe uma associação linear negativa, ou seja, ao passo que uma variável aumenta a outra variável diminui. Finalmente, na Figura 10.5(c)

as variáveis não tem nenhuma associação linear, ou seja, quando uma variável aumenta a outra variável não aumenta e nem diminui linearmente.



Quanto mais próximos de uma reta os pontos do diagrama de dispersão estiverem mais forte será a correlação entre as duas variáveis e quanto mais dispersos os pontos estiverem mais fraca será a correlação entre as duas variáveis. Pode-se determinar o grau de correlação entre duas variáveis utilizando-se, para isso, uma medida.

10.2.2 Coeficiente de correlação

Antes de se definir o coeficiente de correlação deve-se, primeiro, definir a covariância. A covariância é uma medida do grau de interdependência de duas variáveis quantitativas e é calculada pela Equação 10.4.

$$cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (10.4)$$

A covariância é uma medida diretamente afetada pela escala de mensuração das variáveis e pode variar entre $-\infty$ e $+\infty$, ou seja, $-\infty \leq cov(X, Y) \leq +\infty$.

Uma outra maneira de avaliar o grau de associação entre duas variáveis quantitativas é padronizando a covariância, dividindo esta medida pelo produto dos desvios padrões das variáveis sob análise, denominada de coeficiente de correlação. O coeficiente de correlação (linear) é uma medida do grau de associação (correlação) entre duas variáveis quantitativas e não é afetado pela escala de mensuração das variáveis.

Definição

Dados n pares de valores $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ de duas variáveis quantitativas X e Y , define-se o coeficiente de correlação entre X e Y pela Equação 10.5.

$$r = cor(X, Y) = \frac{cov(X, Y)}{dp(X).dp(Y)} \quad (10.5)$$

em que $dp(X)$ é o desvio padrão de X e $dp(Y)$ é o desvio padrão de Y . O coeficiente de correlação pode variar de -1 a 1 , ou seja: $-1 \leq r \leq 1$.

Um coeficiente de correlação positivo indica uma associação linear positiva entre as duas variáveis, um coeficiente de correlação negativo indica uma associação linear negativa, já um coeficiente de correlação nulo (igual a zero) indica que não existe associação linear entre as duas variáveis. Quanto mais próximo de 1 o coeficiente de correlação estiver, mais forte é o grau de associação linear positiva entre X e Y , e, quanto mais próximo de -1 o coeficiente de correlação estiver, mais forte é o grau de associação linear negativa entre X e Y .

A Equação 10.5 pode ser operacionalizada de modo mais conveniente pela Equação 10.6.

$$r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) \left(\sum_{i=1}^n y_i^2 - n \bar{y}^2 \right)}} \quad (10.6)$$

■ **Exemplo 10.7** Considere os dados apresentados no Exemplo 10.6 (Tabela 10.9). Para calcular o coeficiente de correlação construir uma tabela de cálculos auxiliares (Tabela 10.10).

Tabela 10.10: Tabela de cálculos auxiliares

ID	(X)	(Y)	(X ²)	(Y ²)	(XY)
1	2	48	4	2304	96
2	3	50	9	2500	150
3	4	56	16	3136	224
4	5	52	25	2704	260
5	4	43	16	1849	172
6	6	60	36	3600	360
7	7	62	49	3844	434
8	8	58	64	3364	464
9	8	64	64	4096	512
10	10	72	100	5184	720
Total	57	565	383	32.581	3.392

Assim, tem-se: $n = 10$, $\bar{x} = 5,7$, $\bar{y} = 56,5$, $\sum x_i^2 = 383$, $\sum y_i^2 = 32.581$ e $\sum x_i y_i = 3.392$.

Logo,

$$\begin{aligned}
 r &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) \left(\sum_{i=1}^n y_i^2 - n \bar{y}^2 \right)}} \\
 &= \frac{3.392 - 10 \times 5,7 \times 56,5}{\sqrt{(383 - 10 \times 5,7^2)(32.581 - 10 \times 56,5^2)}} \\
 &= 0,8768.
 \end{aligned}$$

Como o coeficiente de correlação foi próximo de 1, significa que existe uma forte correlação linear positiva entre o tempo de serviço (X) e o número de clientes (Y) dos agentes da companhia de seguros.

■

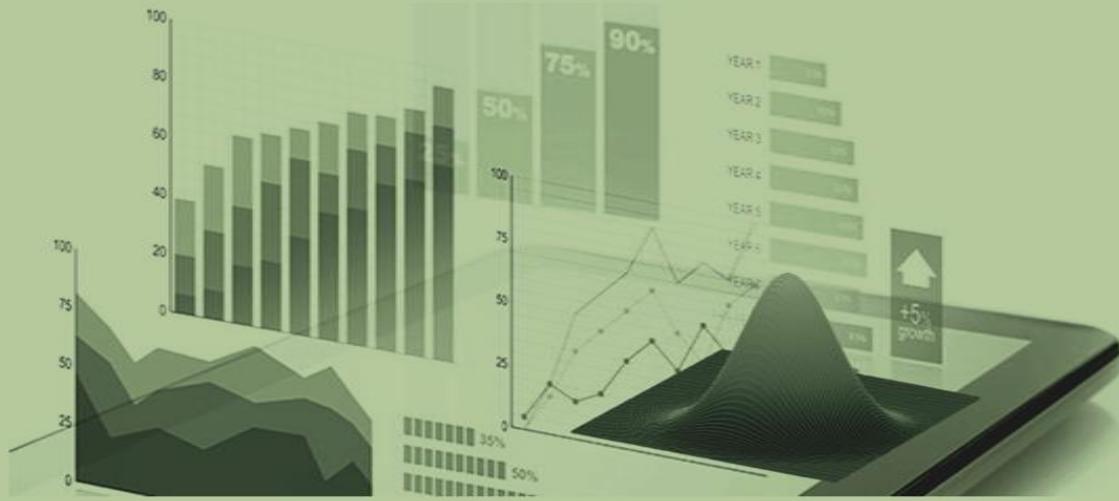
Na Lista 10.5 são apresentados os comandos do software R para obter o gráfico de dispersão X versus Y , a covariância de X e Y e o coeficiente de correlação.

```

1 # Entrando com os dados no R
2 X <- c(2, 3, 4, 5, 4, 6, 7, 8, 8, 10)
3 Y <- c(48, 50, 56, 52, 43, 60, 62, 58, 64, 72)
4
5 # Gráfico de dispersão X versus Y
6 plot(X, Y, pch=19, xlab="X: tempo de serviço, em anos",
7       ylab="Y: número de clientes")
8
9 # Covariância
10 cov(X, Y)
11
12 # Coeficiente de correlação
13 cor(X, Y)

```

Lista 10.5: Comandos do software R



Bibliografia

- BUSSAB, W. O., MORETTIN, P. A. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2002.
- COSTA, S. C. **Estatística Aplicada à Veterinária**. Londrina: UEL, [ca. 2012]. (Notas de aula).
- FERREIRA, D. F. **Estatística Básica**. Lavras: Editora UFLA, 2005.
- FONSECA, J. S., MARTINS, G. A. **Curso de estatística**. 6. ed. São Paulo: Atlas, 1996.
- MELLO, M. P., PETERNELLI, L. A. **Conhecendo o R: uma visão mais que Estatística**. Viçosa: Editora UFV, 2013.
- MONTGOMERY, D. C., RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 5. ed. Rio de Janeiro: LTC, 2012.
- R DEVELOPMENT CORE TEAM (2020). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>.
- RIBEIRO JUNIOR, P. J. **Introdução ao Ambiente Estatístico R**. Curitiba: UFPR, 2011. (Notas de aula).
- RSTUDIO TEAM (2020). **RStudio: Integrated Development for R**. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.
- VIEIRA, S. **Introdução à Bioestatística**. 3. ed. rev. Rio de Janeiro: Campus, 1998.