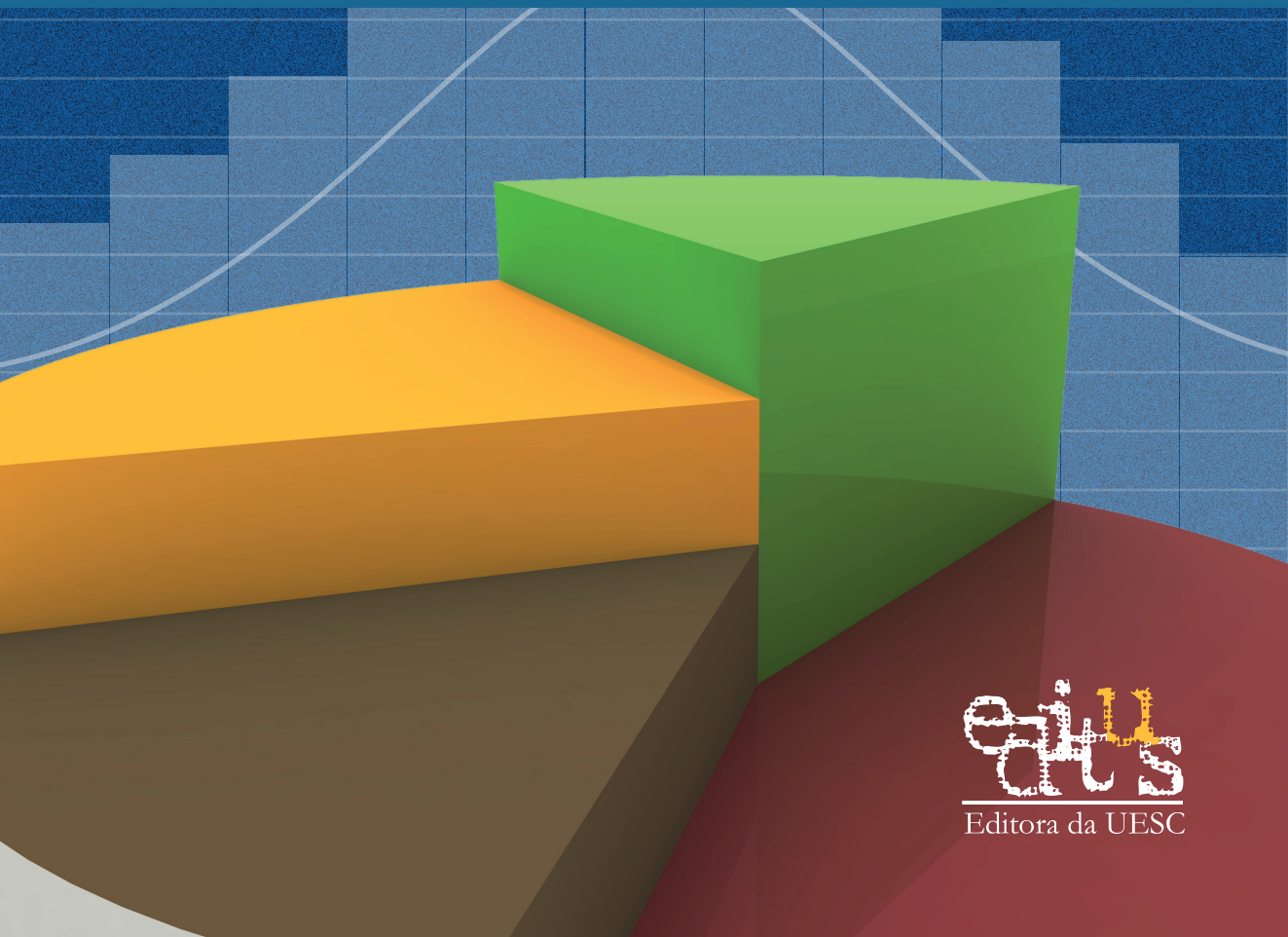
Enio Jelihovschi

Análise Exploratória de Dados usando o R



Editora da UESC

Enio Jelihovschi

Análise Exploratória de Dados usando o R

Ilhéus – Bahia
Editora da UESC
2014

Universidade Estadual de Santa Cruz

Depósito legal na Biblioteca Nacional,
conforme Lei no 10.994, de 14 de dezembro de 2004.
CAPA
Alencar Júnior, com imagem de Jenny Rollo (www.freeimages.com)
REVISÃO
Maria Luiza Nora de Andrade
Dados Internacionais de Catalogação na Publicação (CIP)

J47 Jelihovschi, Enio.
Análise exploratória de dados usando o R/
Enio Jelihovschi– Ilhéus, BA: EDITUS, 2014.
85 p.: il.

Inclui Referências.
ISBN: 978-85-7455-370-2

1. Estatística-Processamento de dados. 2.
R(Linguagem de programação de computador).
Análise multivariada (Processamento de dados
I. Título.

CDD 519.5

EDITUS - EDITORA DA UESC
Universidade Estadual de Santa Cruz
Rodovia Jorge Amado, km 16 - 45662-900
Ilhéus, Bahia, Brasil
Tel.: (73) 3680-5028
www.uesc.br/editora
editus@uesc.br
EDITORA FILIADA À

Associação Brasileira
das Editoras Universitárias

Sumário

Lista de Tabelas

Lista de Figuras

Prefácio

A importância da análise exploratória de dados tem seguido uma trajetória crescente, à medida que o poder de processamento e o tamanho da memória dos computadores foram aumentando, mesmo que os computadores tivessem seu tamanho diminuído. Muitos métodos para Análise Exploratória de Dados foram sendo criados e melhorados, e *softwares* foram sendo escritos à medida em que aquele processo foi tomando força.

Pode-se afirmar que a maior criação estatística dos últimos vinte anos foi, sem dúvida, o Ambiente Computacional Estatístico R, ou somente R como é mais conhecido entre seus usuários; ambiente este que explorou, da forma mais eficiente possível, todo aquele poder computacional, na criação de um *software* único para toda a necessidade e possibilidade computacional de que a estatística necessitava.

Por esta razão, este livro é baseado no uso do R. Optei, porém, por apresentar somente os códigos em R, e não acrescentar um capítulo introdutório com um curso básico de R, isto porque já existem muitos livros com essa abordagem, principalmente na língua inglesa. Em português, posso citar com toda a segurança o excelente livro *R para cientistas sociais*, de Jackson Alves de Aquino, que pode ser baixado gratuitamente do site <http://www.uesc.br/editora/livrosdigitais_20140513/r_cientistas.pdf>, onde um excelente curso básico de R pode ser encontrado. Não poderia fazer melhor do que o Jackson fez. Agradeço-lhe, também, pela gentileza de me haver enviado o código fonte com o qual formatou seu

livro. Ele foi de grande ajuda para a formatação de meu livro, *Análise Exploratória de Dados usando o R*.

Os códigos em R e seus resultados aparecem completos, assim, qualquer leitor do livro, mesmo que seja iniciante em R, e suponho que a grande maioria o seja, poderá copiar os códigos, mudar somente os dados, e rodá-los em R para reproduzi-los com seus próprios resultados. Além disso, os exercícios, e são poucos, são exclusivos para uso do R. Outros exercícios, puramente estatísticos, podem ser encontrados em qualquer livro de Estatística Básica.

Não menos importante são as explicações dos métodos estatísticos na análise exploratória de dados que seguem o meu modo de pensar de como devem ser descritos e ensinados. Bastante ênfase é dedicada à definição de "dados", e às informações que eles contêm. Repetidas vezes, no livro, os dados não são números misteriosos que, repentinamente, aparecem à nossa frente, como num passe de mágica; mas, sim, resultados da medição de uma variável sobre os elementos da amostra, os quais devem ter sido coletados de forma a refletir as nuances da população que queremos estudar. Ou seja, variável e população definem completamente quais informações devem ser alcançadas. Por outro lado, tenho também a propensão e o gosto pela prolixidade descritiva, isto é, gosto das palavras e gosto de fazer uso delas. Quis fazer este livro o mais próximo possível de um livro de estórias.

Capítulo 1

Introdução à Estatística

Antes de começarmos a descrever Estatística como ciência, vamos relatar alguns dos resultados mais importantes alcançados pela ciência no século XX, resultados estes que somente aconteceram por causa do suporte total desta área do conhecimento.

Podemos dizer que o século XX foi o século da Estatística, quando os cientistas se deram conta de que todos os dados medidos nas ciências experimentais têm um resultado aleatório e não pré-determinado, como se pensava anteriormente. Nenhuma outra ciência se desenvolveu tanto nestes últimos 100 anos; pois, partindo do zero, e, num único século, atingiu a importância que tem para a nossa civilização.

Hoje, a população humana no planeta já ultrapassa o número de 6 bilhões de pessoas. Como foi possível produzir comida para alimentar tanta gente? Jamais teria sido possível chegar a este feito sem a revolução que a Estatística criou nos métodos de pesquisa agrícola: como, por exemplo, a produção de sementes de soja adaptadas ao solo e ao clima do cerrado brasileiro. Foi, também, o que aconteceu no caso das vacinas que levaram à erradicação, ou quase erradicação de doenças, como a poliomelite, a varíola, que aleijavam ou matavam milhões de pessoas em todo o mundo. Somente com o uso da Estatística estas vacinas puderam obter a comprovação da sua eficácia.

A Biologia, a ciência da vida, que estuda os animais e os vegetais, e a Ecologia, que estuda sua interação na natureza, nos ensinam como produzir alimento e riqueza sem destruir o meio ambiente. Estas duas áreas de conhecimento usam tanto a Estatística que, a partir delas, criou-se um novo ramo: a Bioestatística, metodologia voltada para a aplicação da Estatística nas ciências da vida.

O que aconteceu foi que, ao longo do século XX, ocorreu uma verdadeira revolução na ciência, representada pela introdução e adoção de métodos estatísticos de pesquisa – que iriam aumentar a confiabilidade das pesquisas na ciência aplicada em geral e, portanto, em seus resultados.

Hoje a Estatística faz parte do nosso dia a dia. Todas as pesquisas de opinião e seus resultados, que escutamos constantemente, são baseadas em seu uso. Quem é que não ouviu falar no IBGE (Instituto Brasileiro de Geografia e Estatística), cujas pesquisas e estatísticas praticamente ordenam a política de investimento e distribuição de renda do governo federal?

A Estatística é a ciência que estuda a forma como toda informação que recebemos por meio de dados pode se tornar inteligível e ser analisada. Quando dizemos informação, falamos daquela básica, usada em pesquisa e análise que, de alguma forma, precisa ser entendida para que possamos utilizá-la. Se escutamos falar que algo novo foi criado, como uma nova semente, um novo aparelho, para que isto chegasse até nós, e pudéssemos nos utilizar do seu resultado, muitos dados que continham informações tiveram de ser classificados, organizados e analisados, tudo muito bem "mastigado"até que a digestão fosse feita e o novo mostrasse a sua utilidade.

A Estatística se dedica à coleta, análise e interpretação dos dados, e para isto algumas das suas práticas são o planejamento, a classificação, a sumarização e a tomada de decisões a partir das observações dos dados. Com isto queremos dizer que a Estatística não é Matemática, ela usa os resultados matemáticos para implementar sua metodologia, e, por isto, não é necessário ser um matemático ou mesmo conhecer alguns

de seus aspectos com profundidade para poder compreender as ideias centrais da Estatística. Para isto, basta aceitar que os dados aparecem de forma totalmente aleatória, sobre os quais não temos quase nenhum controle; em outras palavras, nunca sabemos, de antemão, qual será o resultado final de um experimento, nem quais serão os resultados das nossas observações. Neste aspecto, a variabilidade é um conceito crucial, pois somente a partir do seu estudo e do entendimento profundo do seu conceito, é que chegamos a separar o joio do trigo. Se entendermos isto, entenderemos a Estatística; que se divide em duas partes principais: Estatística Descritiva ou Análise Exploratória de Dados e Inferência Estatística.

Na Estatística Descritiva, usamos métodos como tabelas, gráficos e medidas para tentar entender quais são as estruturas fundamentais dos dados que queremos analisar; e se analisamos dados vindos de fontes ou variáveis diferentes, também tentamos entender as estruturas que relacionam as fontes entre si.

Na Inferência Estatística, fazemos o que se entende por modelagem, ou, melhor dizendo, postulamos um modelo de população de onde provieram os dados e analisamos para concluir se os dados corroboram este modelo ou não.

Neste livro, vamos estudar somente a parte de análise exploratória de dados, isto é, vamos aprender como alguns métodos estatísticos nos ajudam a retirar a informação que aqueles dados estão guardando, no segredo da sua desorganização inicial. Vamos organizar a desorganização de tal maneira que ela irá nos ajudar no nosso entendimento da informação nela contida.

Este livro está baseado no Ambiente Computacional Estatístico R, (RCORETEAM, 2013). Isto quer dizer que todos os exemplos serão escritos na linguagem do R e os resultados serão todos no formato do R. Este livro foi escrito usando a interface gráfica Tinn-R (FARIA; GROSJEAN; JELIHOVSCHI, 2013), feita para facilitar a programação no R, como também permite o uso de outros ambientes computacionais, sendo o Latex um deles. A formatação do livro foi feita usando o Latex, usando

o pacote Knitr do R, (XIE, 2013), que facilita a integração do R com o Latex.

Abaixo temos como exemplo um código em R, que gera 20 números aleatórios, segundo a distribuição normal, a mais importante do cálculo de probabilidades, e calcula a média e a variância desses números. Depois outro código que gera, a partir dos 20 números, um diagrama de caixa (*boxplot*) e histograma, métodos muito usados na análise exploratória de dados.

```
set.seed(1121)
(x ← rnorm(20))
```

```
 [1]  0.14496  0.43832  0.15319  1.08494  1.99954 -0.81188  0.16027
 [8]  0.58589  0.36009 -0.02531  0.15088  0.11008  1.35968 -0.32699
[15] -0.71638  1.80977  0.50840 -0.52746  0.13272 -0.15594
```

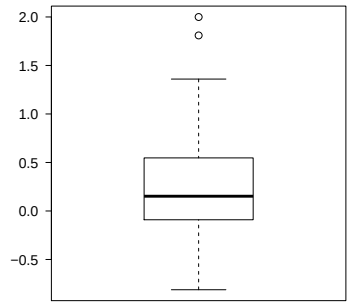
```
mean(x)
```

```
[1] 0.3217
```

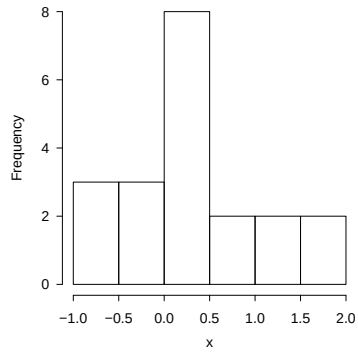
```
var(x)
```

```
[1] 0.5715
```

```
par(mar = c(4, 4, 0.1, 0.1), cex.lab = 0.95, cex.axis = 0.9, mgp
    = c(2,
    0.7, 0), tcl = -0.3, las = 1)
boxplot(x)
```




```
hist(x, main = "")
```



O livro está dividido em duas partes. Na primeira, vamos estudar como explorar estatisticamente dados univariados. Isto quer dizer que estudaremos somente uma variável e o resultado da medida desta variável nos elementos da amostra, conceitos que estudaremos a seguir. Somente no capítulo sobre *correlação* será estudada a relação entre duas variáveis e, neste caso, os dados serão *bivariados*, o que significa que duas variáveis diferentes são medidas no mesmo conjunto de objetos.

Na segunda parte, vamos estudar como explorar estatisticamente dados multivariados. Antes, porém, devo responder à pergunta que tenho certeza já inquieta a mente do leitor: o que são dados multivariados? Os dados multivariados são o resultado da leitura de várias variáveis

no mesmo conjunto de objetos, ou seja, todas variáveis são medidas em cada objeto da amostra selecionada.

As variáveis podem ser categóricas, quantitativas ou uma mistura das duas.

No caso multivariado, além de podermos analisar cada variável em separado, usando os métodos de dados univariados relacionados na primeira parte do livro, também temos de analisar a associação entre estas variáveis, respondendo a perguntas tais como:

- Será que as variáveis são todas independentes entre elas? Se isto acontecer, este seria um caso sem o menor interesse multivariado, isto porque, sendo independentes, estas variáveis podem ser analisadas uma a uma, em separado, e nenhuma informaçao estará perdida.

- Caso contrário, existe uma associação entre as variáveis. Que tipo de relação será esta? Neste caso, toda informação sobre esta associação é de suma importância porque, se os dados foram coletados neste formato (multivariado), é porque, segundo os pesquisadores que coletaram estes dados, muitas informações importantes estão contidas nas associações entre estas variáveis.

- Será que podemos agrupar as variáveis de modo que cada grupo contenha as que estão associadas, sendo as de outro grupo independentes daquelas?

- Será que algumas variáveis explicam uma formulação importante feita a partir dos dados, enquanto outras explicam muito pouco ou praticamente nada?

Parte I

Dados univariados

Capítulo 2

Variáveis

Na linguagem da informação falamos sempre de dados. Para conseguir a informação que queremos ou de que necessitamos, precisamos destes dados. Porém, o que são, na realidade, estes dados?

Na verdade, os dados são o resultado de uma medida: medimos algo. Mas o que vem a ser este algo? Este algo que medimos é o que chamamos de variável. Variável, então, é um ente abstrato que se torna concreto no momento em que a medimos num objeto escolhido para isto. Por exemplo, vou usar a variável peso, usando a unidade de medida de kg. Escolho você que está lendo este livro como objeto de estudo e, ao usar uma balança para medir o seu peso, estou medindo a variável peso e criando um dado. Se em lugar de escolher um leitor, seleciono um grupo de leitores e meço a variável peso em todos, crio, assim, um conjunto de dados. Então, na realidade, temos os dados como a realização da variável que é um ente abstrato e todos os cálculos serão feitos com os dados, entretanto a interpretação dos resultados vai refletir a informação que queremos e é feita voltando às variáveis e seu significado. Se estamos usando a variável peso, então todos os resultados estarão relacionados com o peso dos objetos selecionados.

As variáveis podem ser **categóricas**, também chamadas de **qualitativas**, ou **quantitativas**.

Variáveis categóricas, como o próprio nome diz, exprimem somente uma qualidade do que vem a ser medido, como, por exemplo, a variável sexo, que somente tem duas possibilidades, masculino ou feminino, ou, então, a variável atendimento que pode ser ruim, razoável ou bom. Variáveis quantitativas são aquelas que podem ser quantificadas por um número.

Variáveis categóricas podem ser *nominais* ou *ordinais*. A variável sexo, por exemplo, é uma variável categórica nominal e a variável atendimento é ordinal porque os três níveis de atendimento estão naturalmente ordenados do pior até o melhor.

Variáveis quantitativas podem ser *discretas* ou *contínuas*. Variáveis discretas são o resultado de contagem. Suponhamos que temos uma amostra de 20 pessoas selecionadas por algum processo descrito na seção 2.1. Para cada pessoa perguntamos se ela gosta de música axé, depois contamos quantas pessoas nesta amostra responderam "sim". A variável X que conta o número de pessoas que responderam "sim"pode receber somente os valores $0, 1, 2, \ldots, 20$, e, portanto, é um variável discreta. Variáveis contínuas são o resultado de uma medida numérica, como peso, altura etc.

2.1 Amostragem

População e amostra

Como foi dito acima, para que a variável se realize, temos de escolher um conjunto de objetos no qual medir a variável. Este conjunto é chamado de *amostra*. Para amostrar os objetos temos de selecioná-los de algum conjunto maior de objetos. Este conjunto maior é chamado de *população*.

Vamos aprofundar um pouco estes conceitos. Suponhamos que você seja candidato a prefeito de Ilhéus e encomende uma pesquisa eleitoral com a finalidade de avaliar suas chances de ser eleito. O instituto de pesquisa precisa, então, entrar em contato com os eleitores da cidade para lhes perguntar sobre sua preferência, ou seja, se votarão em você ou não.

Porém sairia demasiado caro para o seu bolso se eles fossem entrevistar todos os eleitores de Ilhéus, cerca de cem mil. Todos os eleitores da cidade constituem a população na qual o instituto está interessado e já vimos que não é viável entrevistar toda a população. Neste caso o instituto vai ter de selecionar uma amostra desta população. Este processo é chamado de amostragem. Esta amostragem tem de seguir um critério estatístico para que o conjunto de objetos selecionados seja representativo da população. Se escolhermos um só bairro da cidade para selecionar a nossa amostra, com toda a certeza a amostra escolhida não será representativa e o resultado obtido não corresponderá à realidade. Você poderia achar que está ganhando, quando, na realidade, no total da cidade você perderia a eleição.

A variável, neste caso, é a questão "em quem você vai votar". Esta é uma variável categórica nominal com dois níveis de resposta. Os dados serão obtidos depois de medir esta variável em cada objeto da amostra. Neste caso, medir significa perguntar. Ao receber a resposta, temos o resultado da medição.

População pode, então, ser definida como o *conjunto total de objetos para os quais quero obter informação*. Amostra, neste caso, é um *subconjunto da população*. Em geral, um subconjunto muito menor do que a população. O mais importante de tudo é que esta amostra seja escolhida de acordo com o processo estatístico chamado *amostragem*.

Amostragem

Existe somente um fenômeno que consegue eximir uma amostra de qualquer arbitrariedade, fazendo com que a amostra reflita, em si, se não todos, pelo menos parte dos atributos da população. Este fenômeno se chama aleatoriedade. Se nosso intuito é fazer um estudo sobre o peso dos estudantes da UESC (Universidade Estadual de Santa Cruz), é claro que não vamos pesar todos os estudantes, mas queremos que nossa amostra reflita bem toda a distribuição de peso dos estudantes, e não somente alguns atletas ou alguns muito gordos. A melhor forma de escolher esta

amostra de, vamos dizer, 100 estudantes, seria sorteá-los de forma que todos os objetos da população tivessem a mesma chance de estar na amostra.

Poderíamos, por exemplo, sortear de uma urna que contivesse todos os nomes, ou usar um gerador de números aleatórios de um computador e sortear os estudantes pelo número de matrícula.

Esta forma de amostragem é chamada de *amostra casual simples*.

Este tipo de amostragem é o melhor que existe. Todas as vezes em que ele puder ser realizado, com certeza a amostra vai espelhar a população. O difícil, porém, é realizar uma amostragem casual simples no mundo real. Imaginem a dificuldade de um instituto de pesquisa sortear 1000 eleitores de um estado como a Bahia e, depois, encontrar todas estas pessoas nos mais diferentes recantos do estado. É realmente muito difícil e muito caro de ser realizado. Portanto, outros esquemas de amostragem são usados.

O primeiro que vamos definir chama-se *amostragem sistemática*. Este tipo de amostragem é usado quando os dados possuem uma ordenação natural como, por exemplo, fichários, prontuários, casas numa rua. Ela é feita da seguinte forma: suponhamos que temos um fichário com 10000 fichas e queremos amostrar 100 fichas. Primeiramente, sorteamos um número entre 1 e 100. Esta é a parte aleatória da amostragem. Suponhamos agora que o número sorteado foi 43, então escolhemos a ficha que está na ordem 43; depois 143, 243 e assim por diante. Desta forma escolhemos as 100 fichas. O importante é que sempre tem de haver uma parte aleatória no processo de amostragem.

O segundo é o processo de *amostragem estratificada*. Muitas vezes uma população é composta de subpopulações (estratos) bem definidas, de forma tal que, usando estes estratos, facilitamos o processo de amostragem. Por exemplo, suponhamos que o instituto de pesquisa quer selecionar 2000 eleitores em todo o estado da Bahia. Como o estado é naturalmente dividido em municípios (estratos), podemos sortear 20 municípios dos mais de 400 em todo o estado. Cada sede de município é naturalmente dividida em bairros (subestratos) então, sorteamos

5 bairros em cada sede. Cada bairro é naturalmente dividido em ruas (sub-subestratos), sorteamos, então, 5 ruas em cada bairro selecionado e as residências, nas ruas, podem ser amostradas segundo o método sistemático descrito acima.

O R possui algumas funções que permitem fazer uma simulação de amostragem. No nosso caso, vamos usar somente a função *sample*(`x, size, replace = FALSE`). "x"é um vetor que pode ser numérico como também vetor de caracteres. Este vetor representa a população. “size” é um número inteiro que representa o tamanho da amostra, ou seja, quantos elementos da população vamos escolher. “replace” representa o tipo de escolha ou sorteio. *Replace* significa repor, se igual a TRUE, então a escolha é feita com reposição do elemento sorteado; caso contrário, sem reposição do elemento.

Abaixo um exemplo de uma amostragem simples. Vamos selecionar 10 letras das 26 letras do alfabeto. Primeiro, sem reposição e, depois, com reposição.

```
set.seed(1121)
sample(letters, size = 10)
```

```
 [1] "o" "g" "q" "h" "m" "u" "r" "i" "t" "f"
```

```
sample(letters, size = 10, replace = TRUE)
```

```
 [1] "f" "m" "o" "w" "s" "h" "q" "y" "m" "h"
```

Note que, no segundo resultado, a letra "h"foi selecionada duas vezes, já que o sorteio foi feito com reposição.

Agora vamos mostrar como simular o processo de *amostragem estratificada* descrito acima. Vamos supor que os municípios, bairros e ruas estão numerados.

```
set.seed(1121)
```

```
munic ← sample(1:400, size = 20)  #número de cada município
    selecionado
sort(munic)  # os municípios números em ordem crescente
```

```
 [1]  82 111 112 113 127 128 169 186 188 219 223 224 246 267 279 339
[17] 340 369 384 385
```

```
municib ← sample(50:150, size = 20)  #número de bairros em cada
     município
sort(municib)  #número de bairros por município em ordem
    crescente
```

```
 [1]  60  61  72  76  77  83  85  86  89  91  96 103 106 108 111 130
[17] 137 138 140 150
```

```
bair ← apply(as.matrix(municib[order(munic)]), 1, function(x)
    sample(x,
     size = 5))
rownames(bair) ← paste("bairro", 1:5)
colnames(bair) ← paste(municib[order(munic)], sort(munic))
bair  #matriz cujas linhas são os números dos bairros sorteados
    de cada município
```

```
         137 82 83 111 130 112 60 113 89 127 140 128 61 169 91 186
bairro 1     73      2      17     32     17     103     49     62
bairro 2     76     51      23     13     44      13     46     52
bairro 3     64     21      85     48     10      72     19     15
bairro 4     40     81      72     31      6      47     32     51
bairro 5     72     42      41     20      2      87     53     79
         86 188 111 219 138 223 106 224 96 246 103 267 76 279 77 339
bairro 1      8     104      44      13     91       5     40     31
bairro 2     65      58     110      85     73      60     19     22
bairro 3     68       4       7       2     22     100     43     42
bairro 4      1      25      42      28     85      17     64      4
bairro 5      7      13     116      62     60      47     21     52
         85 340 108 369 72 384 150 385
bairro 1     70     101     49      55
bairro 2      8      98     50       6
bairro 3     27      37     56     131
bairro 4     75      61     63      80
```

```
bairro 5      3      47     48      86
```

Na matriz acima, podemos ver, na primeira, coluna que o município 82 possui 137 bairros e, destes bairros, foram escolhidos os bairros 73, 76, 64, 40, 72. A segunda coluna mostra que o município 111 possui 83 bairros, sendo que os bairros 2, 51, 21, 81, e 42 foram os escolhidos, e assim por diante.

2.2 Exercícios

1. Um auditor precisa selecionar uma amostra de 120 empréstimos para auditar empréstimos realizados pelo banco onde trabalha. Ele possui um arquivo com 5000 fichas de empréstimos, organizadas em ordem alfabética, no fichário do banco. Escreva um código, no R, que simule este processo de amostragem, e colete a amostra.

2. Um instituto de pesquisa foi contratado para fazer uma pesquisa eleitoral com a finalidade de avaliar a proporção do eleitorado de um estado que deverá votar no candidato contratante na próxima eleição para governador. O estado citado possui 88 municípios. Escreva um código no R que simule o sistema de *amostragem estratificada*, selecionando 10 municípios, 10 bairros por município, 10 ruas por bairro e as casas selecionadas, segundo o sistema sistemático.

Capítulo 3

Tabelas

3.1 Tabela de variáveis categóricas

Tabela é uma forma de organizar os dados. Em geral, é o primeiro passo a ser dado com a finalidade de tornar as palavras, letras ou números, ou seja, os nossos dados, mais informativos. Imagine se você coletasse uma amostra de uma variável categórica, um questionário cujas respostas fossem *ótimo, bom, regular, ruim e péssimo*. O tamanho da amostra é de 100 objetos. Neste caso, os dados brutos são duzentas repetições destas palavras. O código no R abaixo mostra o resultado desta amostra simulada.

Olhe para ela. Será que você consegue obter alguma informação útil destes dados postos desta forma? É difícil, não é?

```
set.seed(1121)
resposta <- sample(c("ótimo", "bom", "regular", "ruim", "péssimo
    "), size = 100,
    replace = TRUE)
resposta
```

```
  [1] "regular" "bom"     "ruim"    "bom"     "regular" "péssimo"
  [7] "péssimo" "regular" "péssimo" "bom"     "bom"     "regular"
 [13] "regular" "péssimo" "ruim"    "bom"     "ruim"    "péssimo"
```

```
 [19] "regular" "bom"     "regular" "bom"     "regular" "regular"
 [25] "péssimo" "regular" "bom"     "ótimo"   "bom"     "péssimo"
 [31] "péssimo" "regular" "ruim"    "bom"     "bom"     "ótimo"
 [37] "regular" "ruim"    "regular" "péssimo" "regular" "regular"
 [43] "regular" "bom"     "regular" "ótimo"   "ruim"    "bom"
 [49] "bom"     "regular" "ótimo"   "ótimo"   "ruim"    "regular"
 [55] "bom"     "regular" "bom"     "péssimo" "regular" "bom"
 [61] "ótimo"   "regular" "ótimo"   "ótimo"   "ótimo"   "ruim"
 [67] "ótimo"   "regular" "bom"     "ruim"    "ruim"    "ruim"
 [73] "bom"     "regular" "péssimo" "ruim"    "regular" "ótimo"
 [79] "regular" "péssimo" "ótimo"   "ruim"    "péssimo" "ótimo"
 [85] "ótimo"   "péssimo" "regular" "ótimo"   "bom"     "ótimo"
 [91] "bom"     "ruim"    "ótimo"   "bom"     "péssimo" "ótimo"
 [97] "péssimo" "ótimo"   "bom"     "ruim"
```

Exatamente por isso vemos a necessidade de organizar estes dados de forma tal que possamos começar a retirar deles a informação útil que queremos. O primeiro método que vamos estudar são as representações tabulares ou, como são chamadas mais comumente, tabelas.

O ato de tabulação, ou seja, fazer uma tabela significa agregar as medidas da variável em questão que tenham uma propriedade em comum. Quando a variável é qualitativa ou quantitativa discreta, esta propriedade é: "serem iguais"(isto quer dizer que juntamos todas as medidas do mesmo valor todos os resultados iguais, e contamos quantas são). Repetimos este procedimento para cada valor da variável. No caso da variável quantitativa contínua, veremos, mais abaixo, como é feita a tabela.

Uma tabela é formada por: título, cabeçalho, coluna indicadora, resultados, fonte. No Brasil, a apresentação e formatação de uma tabela é regida pela Associação Brasileira de Normas Técnicas (ABNT).

Abaixo está o código no R para fazer uma tabela. O pacote **xtable**, (DAHL, 2013) é usado para isto. Este pacote transforma uma tabela do R, usada para cálculos, nesta tabela escrita em Latex.

```
library(xtable)
frequen ← table(resposta)[c(2, 1, 4, 5, 3)]
```

```
frequencia ← data.frame(names(frequen), frequen)
rownames(frequencia) ← NULL
colnames(frequencia) ← c("resposta", "freq")

print(xtable(frequencia, caption = "Questionário, ano 2012,
    Ilhéus", label = "tab:tabelas2"),
    table.placement = "H", caption.placement = "top",
        latex.environments = "flushleft",
    include.rownames = FALSE, hline.after = c(-1, 0), add.to.row
        = list(pos = list(5),
        command = c("\\hline\\multicolumn{2}{p{4cm}}\n{\\
            footnotesize\nFonte: Dados hipotéticos.}")))
```

Tabela 3.1: Questionário, ano 2012, Ilhéus

resposta	freq
ótimo	19
bom	23
regular	27
ruim	15
péssimo	16

Fonte: Dados hipotéticos.

O título da tabela mostra "o que, quando e onde"o estudo, que gerou a tabela, foi realizado e seu nome. No cabeçalho, estão os nomes das duas colunas que mostram o que representa o estudo, "resposta e frequência". A coluna indicadora mostra o significado, ou nível, de cada resultado, “ótimo, bom, regular, ruim, péssimo". E, finalmente, temos o corpo da tabela, onde se encontram as frequências tabuladas para cada nível da coluna indicadora.

3.2 Tabela de distribuição de frequências

Quando a variável usada é quantitava e contínua, não é possível tabular o resultado de sua medida nos objetos da amostra, pois os dados gerados não terão valores repetidos. Repetições de valores acontecem somente quando arredondamos os resultados, como, por exemplo, nas medidas de peso de pessoas. Nunca dizemos que pesamos 67,34 kg, medimos somente a parte inteira da medida, porém, se o instrumento de medição for bastante preciso, duas pessoas jamais terão exatamente o mesmo peso. Neste caso, é fácil ver que não podemos contar resultados iguais, como foi feito no processo de tabulação descrito acima. Como fazer, então?

O que se faz, neste caso, é juntar dados com valores que estejam "perto"uns dos outros, o que chamamos intervalos de classe. A frequência de cada intervalo será o número de dados cujos valores pertencem a este intervalo.

O formato da tabela de distribuição de frequência é o mesmo da outra tabela já descrita, contendo o título, cabeçalho, coluna indicadora, resultados e a fonte. O pacote do R, chamado fdth, (FARIA; JELIHOVSCHI, 2012), é usado para fazer estas tabelas.

Tabela 3.2: Tabela de distribuição de frequências

intervalo de classe	f	fr	fr%	d=fr/int
[1, 2)	0	0	0	0
[2, 3)	1	0.02	2	0.02
[3, 4)	3	0.06	6	0.06
[4, 5)	20	0.40	40	0.40
[5, 6)	21	0.42	42	0.42
[6, 7)	5	0.1	10	0.1

Fonte: Dados hipotéticos.

Olhe a tabela 3.2. Será que podemos entender o porquê do nome "tabela de ditribuição de frequência"? No fundo, ela nos mostra a frequência

com que cada intervalo aparece na nossa amostra de dados, ou seja, quantos valores dos dados estão em cada intervalo. Neste caso, f (frequência absoluta) é o resultado da contagem de dados em cada intervalo, fr (frequência relativa) é f/(total de dados), fr% (frequência relativa em percentual) e, finalmente, d (densidade) é igual a fr/(tamanho do respectivo intervalo de classe). Isto quer dizer que ela mostra qual é a distribuição das frequências dos intervalos de classe.

Voltando ao nosso exemplo das medidas dos pesos, nosso interesse é saber como os pesos das pessoas pertencentes à população estudada se distribuem ao longo de um dado intervalo. Quais são os intervalos que contêm a maioria dos pesos? Será que os pesos poderiam ser modelados por uma curva de distribuição conhecida? Neste caso, esta tabela poderá fornecer muitas informações, como: qual intervalo de pesos poderá ser considerado “normal"na nossa população? a partir de qual peso uma pessoa será considerada obesa ou muito magra? Então, queremos uma tabela que seja o mais informativa possível sobre esta distribuição de pesos da população. Para isto, o mais importante será encontrar o número ótimo de intervalos de classe e, por consequência, o seu tamanho. Se usamos poucos intervalos, as frequências ficam muito altas em cada um deles e não obtemos muita informação, como na tabela abaixo. Não obtemos muita informação vendo que o intervalo de 0.5 até 4.5 contém 15 dados, e o intervalo de 4.5 até 8.5 contém 35 dados.

```
library(fdth)
set.seed(3051952)
x ← rnorm(n = 50, mean = 5, sd = 2)
x
```

```
 [1] 3.0991 3.2133 5.6626 4.3937 5.4752 6.1914 6.5227 3.8266 7.6269
[10] 6.0379 2.8208 6.7012 7.5582 7.7148 0.7556 5.5554 7.3544 4.7410
[19] 9.4985 4.1056 8.1968 3.9699 6.8870 4.9730 4.2042 3.5618 8.5532
[28] 3.9393 5.4301 3.0582 9.2737 8.1115 5.8928 6.4750 3.7498 4.4239
[37] 5.0554 2.4379 3.5870 6.8041 4.6939 1.6521 3.3523 3.9483 5.4860
[46] 5.8611 8.3815 1.1665 7.3435 7.3694
```

```
d ← fdt(x, start = 0.5, end = 8.5, h = 4)
print(d, format = TRUE, col = 1:4, pattern = "%.2f")
```

```
Class limits  f   rf rf(%)
[0.50, 4.50) 20 0.40    40
[4.50, 8.50) 27 0.54    54
```

Por outro lado, se usamos muitos intervalos, as frequências dos intervalos ficarão muito baixas e, neste caso, não obtemos quase nenhuma informação sobre a real distribuição de frequência dos dados.

```
d ← fdt(x, start = 0.5, end = 8.5, h = 0.5)
print(d, format = TRUE, col = 1:4, pattern = "%.2f")
```

```
Class limits f   rf rf(%)
[0.50, 1.00) 1 0.02     2
[1.00, 1.50) 1 0.02     2
[1.50, 2.00) 1 0.02     2
[2.00, 2.50) 1 0.02     2
[2.50, 3.00) 1 0.02     2
[3.00, 3.50) 4 0.08     8
[3.50, 4.00) 7 0.14    14
[4.00, 4.50) 4 0.08     8
[4.50, 5.00) 3 0.06     6
[5.00, 5.50) 4 0.08     8
[5.50, 6.00) 4 0.08     8
[6.00, 6.50) 3 0.06     6
[6.50, 7.00) 4 0.08     8
[7.00, 7.50) 3 0.06     6
[7.50, 8.00) 3 0.06     6
[8.00, 8.50) 3 0.06     6
```

Existem alguns métodos desenvolvidos por estatísticos que dão uma fórmula para o número ótimo de intervalo, a depender da quantidade de dados, ou seja, do tamanho da amostra. Os métodos mais conhecidos são os de Sturges, de Scott e de Freedman-Diaconis (FD). O método usado por omissão é o de Sturges. Todos três são comumente usados. Dependendo dos dados e da população de onde eles são coletados, um tem preferência sobre o outro.

No método de Sturges, se n é o tamanho da amostra, então o número de intervalos de classe será $1 + 3.22 * log(n)$, sendo “log"o logaritmo na base 10. O pacote do R chamado **fdth** facilita muito na feitura de uma tabela no R. No caso acima, foram gerados dados aleatórios contidos no vetor "x".

*Como usar o **fdth***

Primeiramente, temos de criar um vetor x no R, onde estão contidos os dados. Depois rodamos a função do pacote chamada **fdt**. Veja abaixo.

```
d ← fdt(x)
print(d)
```

```
Class limits  f   rf rf(%) cf cf(%)
[0.748,2.01)  3 0.06     6  3     6
 [2.01,3.28)  5 0.10    10  8    16
 [3.28,4.54) 12 0.24    24 20    40
  [4.54,5.8)  9 0.18    18 29    58
  [5.8,7.07)  9 0.18    18 38    76
 [7.07,8.33)  8 0.16    16 46    92
 [8.33,9.59)  4 0.08     8 50   100
```

Como podemos ver, usando o método de Sturges, o número de intervalos foi 7, cada um com tamanho de 1, 1. Então sabemos que o melhor número de intervalos é 7 e podemos ajustar o tamanho para que fique mais fácil captar a ideia destes intervalos.

```
d ← fdt(x, start = 0.5, end = 8.5, h = 1)
print(d, format = TRUE, col = 1:4, pattern = "%.2f")
```

```
Class limits  f   rf rf(%)
[0.50, 1.50)  2 0.04     4
[1.50, 2.50)  2 0.04     4
[2.50, 3.50)  5 0.10    10
[3.50, 4.50) 11 0.22    22
```

```
[4.50, 5.50)  7 0.14    14
[5.50, 6.50)  7 0.14    14
[6.50, 7.50)  7 0.14    14
[7.50, 8.50)  6 0.12    12
```

Usando os argumentos da função fdt e da função *print* acima, vemos o resultado formatado como interessa mostrar, ou seja, os intervalos de classe (*class limits*), frequência absoluta (f), frequência relativa (rf) e frequência relativa em percentual. A densidade não é mostrada na tabela da função fdt.

3.3 Tabela de contingência

Tabela de contingência é uma tabela de tabulação cruzada de duas ou mais variáveis categóricas, podendo também ser chamada de fatores. Cada variável pode ser dicotômica, ou seja, ela somente tem dois resultados, também chamados de níveis; ou politômica, com vários níveis.

Por exemplo, a variável sexo é dicotômica, enquanto a variável classe social é politômica, classe A, B, C ou D.

O código abaixo mostra como construir uma tabela de contingência.

```
# Gerando os dados: variável sexo
set.seed(3051952)
sexo ← sample(c("F", "M"), size = 75, replace = TRUE)
sexo
```

```
 [1] "F" "F" "F" "F" "M" "M" "F" "M" "M" "F" "M" "M" "M" "F" "F" "F"
[17] "M" "F" "M" "F" "F" "M" "M" "F" "M" "F" "M" "F" "F" "F" "M" "M"
[33] "M" "M" "F" "F" "M" "F" "F" "M" "M" "F" "F" "F" "M" "M" "F" "M"
[49] "F" "M" "F" "M" "M" "M" "F" "F" "M" "F" "F" "F" "M" "M" "M" "M"
[65] "M" "M" "M" "M" "F" "M" "F" "F" "M" "M" "F"
```

```
# Gerando os dados: variável classe social (clso)
clso ← sample(LETTERS[1:4], size = 75, replace = TRUE)
clso
```

```
 [1] "B" "A" "B" "D" "C" "B" "D" "A" "B" "A" "D" "B" "C" "C" "B" "C"
[17] "C" "D" "C" "A" "C" "D" "C" "D" "A" "A" "C" "D" "B" "D" "B" "B"
[33] "D" "D" "D" "D" "A" "A" "D" "D" "B" "A" "D" "A" "C" "D" "B" "A"
[49] "D" "D" "C" "D" "B" "B" "D" "A" "B" "C" "C" "B" "D" "C" "C" "A"
[65] "C" "C" "B" "B" "C" "A" "D" "C" "D" "B" "C"
```

```
# fazendo a tabela
tab ← table(sexo, clso)
tab
```

```
    clso
sexo  A  B  C  D
   F  8  6  9 13
   M  6 12 11 10
```

```
# Usando o pacote xtable para construir uma tabela em Latex
print(xtable(tab, caption = "tabela de contingência", label = "
    tab:tabelas7a"),
    table.placement = "H", caption.placement = "top",
        latex.environments = "flushleft",
    hline.after = c(-1, 0), add.to.row = list(pos = list(2),
        command = c("\\hline\\multicolumn{5}{p{4cm}}\n{\\Tiny\
        nFonte: Dados hipotéticos.}")))
```

Tabela 3.3: tabela de contingência

	A	B	C	D
F	8	6	9	13
M	6	12	11	10

Fonte: Dados hipotéticos.

3.4 Exercícios

1. Usando a função *sample*, gere dados aleatórios a partir do vetor 4000:7000 (população), com 50 dados, sem reposição. Use a função fdt para fazer uma tabela de distribuição de frequências dos dados.

2. Repita o mesmo para 100 dados.

3. Repita para 300, 500 e 1000 dados.

4. No pacote do R *datasets*, chame o conjunto de dados *EuStockMarkets* e use o pacote *fdth* para fazer uma tabela de distribuição de frequências para cada uma das variáveis. Preste atenção que *EuStockMarkets* é um *data.frame*.

Capítulo 4

Visualização gráfica de dados

A parte do cérebro humano destinada à codificação da visão é maior do que as partes destinadas aos outros quatro sentidos juntos. Isto significa que o ser humano utiliza a visão muito mais que os outros sentidos. Tanto assim que, muitas vezes, usamos o verbo ver no sentido de ouvir, de degustar, de cheirar ou mesmo de tatear alguma coisa.

Por esta razão, uma apresentação gráfica de números ou relações numéricas nos parecem muito mais informativas do que somente o uso de tabelas. Neste capítulo, veremos como transformar tabelas em gráficos, assim como a forma de usá-los e interpretá-los.

Da mesma forma que as tabelas, os gráficos têm títulos e fonte do fornecedor das informações que geraram o gráfico.

4.1 Gráfico de colunas

O gráfico de colunas é assim chamado por ser feito de colunas cujas alturas representam os valores das respectivas frequências de uma tabela de variável categórica.

A seguir, o gráfico de colunas da tabela 3.1.

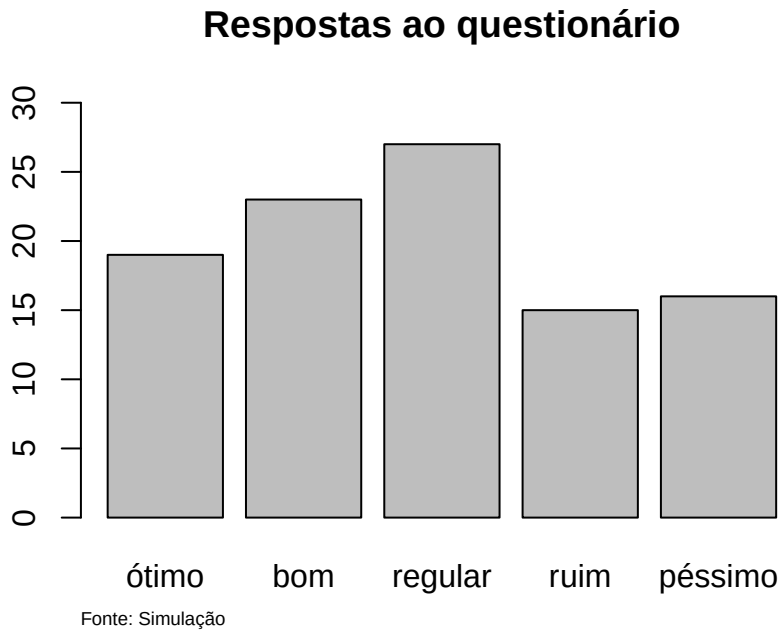


Figura 4.1: Gráfico de coluna

```
set.seed(1121)
resposta ← sample(c("ótimo", "bom", "regular", "ruim", "péssimo
    "), size = 100,
    replace = TRUE)
frequen ← table(resposta)[c(2, 1, 4, 5, 3)]
```

```
barplot(frequen, ylim = c(0, 30), main = "Respostas ao
    questionário")
mtext("Fonte: Simulação", side = 1, line = 2, adj = 0, cex = 0.6
    )
```

É importante prestar atenção na forma como é feita a marcação no eixo vertical.

1. Antes de tudo, marcamos a régua proporcional de acordo com a frequência máxima. Por exemplo, a tabela 3.1 tem a frequência maxima de 27, portanto o valor máximo a ser marcado no eixo vertical será 30.

2. Dividimos o intervalo, no eixo, de 5 em 5 até chegar em 30.

3. Marcamos o valor de cada frequência nesta régua, da mesma forma que medimos usando uma régua de plástico marcada em centímetros.

Ao compararmos o gráfico 4.1 com a tabela 3.1, a informação a respeito das relações entre as frequências, isto é, a ideia do valor de cada uma, como também no quanto cada uma é proporcionalmente diferente das outras, salta aos olhos. Num gráfico, assimilamos as informações muito mais rapidamente do que observando somente a tabela que deu origem àquele gráfico.

A apresentação de um gráfico tem de ter um título, a fonte de onde foram coletados os dados e os nomes da variável representada no(s) eixo(s).

4.2 Gráfico de setores

Gráfico de setores, também conhecido como gráfico de pizza, é um gráfico especial, para variáveis categóricas, cuja informação mais importante está numa tabela, na qual a frequência, calculada em cada nível da variável, é do tipo relativa, escrita em números de 0 a 1 ou em percentagens. Os dados são hipotéticos.

```
partes ← c(10, 12, 4, 16, 8)
paises ← c("EUA", "Inglaterra", "Australia", "Alemanha", "
    França")
pct ← round(partes/sum(partes) * 100)
```

```
paises <- paste(paises, pct)  # agregando percentagens aos nomes
paises <- paste(paises, "%", sep = "")  # adicionando o sinal %
pie(partes, labels = paises, col = rainbow(length(paises)), main
    = "Gráfico de setores de paises.")
```

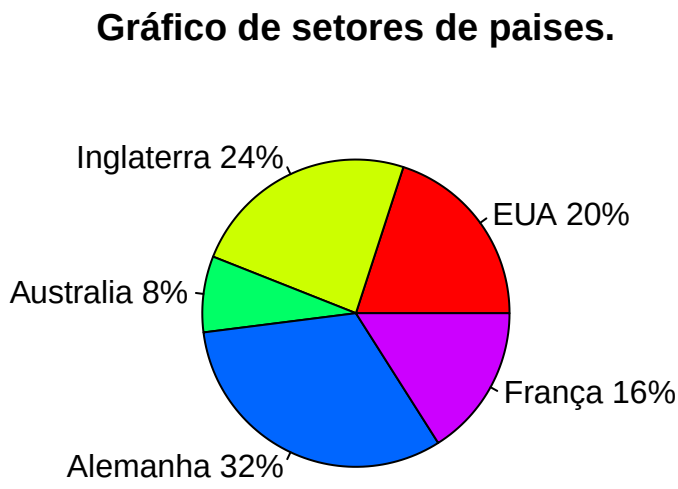


Figura 4.2: Gráfico de setores

4.3 Histograma

O histograma é um gráfico de colunas feito a partir de uma tabela de distribuição de frequências, com a particularidade de não haver es-

paço entre as colunas. O pacote *fdth* também é usado para fazer os histogramas.

A seguir está o código no R, usando o pacote *fdth*, para se fazer um histograma, e o resultado plotado abaixo.

```
library(fdth)
set.seed(1)
x ← rnorm(n = 50, mean = 5, sd = 2)
d ← fdt(x, start = 0.5, end = 8.5, h = 1)
plot(d)
```

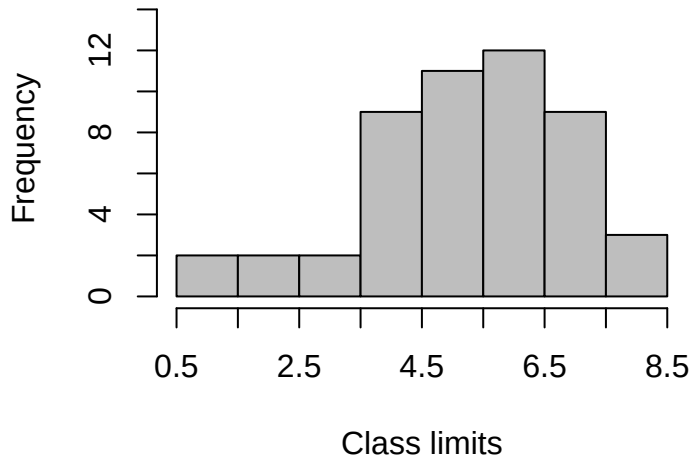


Figura 4.3: Histograma

4.4 Exercícios

1. Faça o histograma da tabela do exercício 1 do capítulo anterior.

2. Faça o histograma da tabela do exercício 2 do capítulo anterior.

3. Faça o histograma das tabelas do exercício 3 do capítulo anterior.

4. Faça o histograma das tabelas do exercício 4 do capítulo anterior. Repita os histogramas, mudando o número de intervalos de classe, tanto para mais como para menos, e veja como os gráficos mudam. Compare com os números de intervalos de classe dados no pacote *fdth*. Leia com atenção a seção *Tabela de distribuição de frequências* do capítulo anterior.

Capítulo 5

Medidas de tendência central

As medidas de tendência central são os resultados de certas operações feitas com os dados que nos dão uma ideia representativa da ordem de grandeza dos valores da variável medida, ou do valor que melhor representa os dados coletados.

5.1 Média

A média de uma série de números é a soma destes números dividida pela quantidade de números que compõe a série. Assim podemos ver que média somente pode ser calculada para dados que resultam da medida de uma variável quantitativa. Ela também pode ser enxergada da seguinte forma: suponhamos que os valores dos nossos dados são marcados sobre uma fina régua de madeira a partir de um ponto, o qual chamamos de origem ou zero. Usando este ponto como referência, marcamos as distâncias correspondentes aos valores dos nossos dados. Se o valor for positivo, marcamos a distância à direita da origem; se for negativo, à esquerda. Em cada ponto marcado, colocamos um peso, o mesmo peso para todos. O ponto cuja distância da origem for a média calculada, como indicado acima, será o centro de gravidade dos pesos sobre a régua. Isto quer dizer que se apoiamos a régua neste ponto, ela

ficará perfeitamente equilibrada, ou seja, não tenderá a cair para nenhum dos lados.

Seja X a variável em questão, peso. Seja x_i para $i = 1 \ldots n$ os valores da variável medida em cada objeto da amostra, ou seja, o peso de cada pessoa pertencente à amostra, e simbolizamos a média desta variável por $\bar{X}$.

A fórmula da média pode então ser indicada por:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

A função do R que calcula a média se chama *mean*. No *R*, calculamos conforme o programa abaixo:

```
set.seed(31051952)
x ← sample(1:1000, size = 20, replace = TRUE)
x
```

```
 [1] 357 237 725  94 389 154 738 869  94 432 674 624 532 197 630 739
[17] 284 176  60 127
```

```
mean(x)
```

```
[1] 406.6
```

Na verdade, a média de todos os números inteiros de 1 até 1000 é 500, no entanto, como foram sorteados somente 20 números, a média será um número perto de 500; mas, muito raramente, poderia ser igual a 500. Se o tamanho da amostra for maior, o valor da média, calculada a partir desta amostra, estará mais perto de 500 do que da anterior.

```
set.seed(31051952)
x ← sample(1:1000, size = 100, replace = TRUE)
mean(x)
```

```
[1] 476.6
```

Podemos, por este motivo, entender melhor o significado da média. Se a média é o ponto de equilíbrio, isto quer dizer que ela representa com fidelidade um resumo da grandeza dos dados: que, neste caso, são as distâncias até a origem, cujos tamanhos são os resultados das medidas da variável quantitativa sobre os objetos da amostra. Por outro lado, é exatamente por isto que a média sofre de um ponto fraco.

Suponhamos que temos uma régua muito comprida. A princípio, agrupamos nossos pesos em pontos não muito longe da origem e calculamos a média: o ponto de equilíbrio. O que aconteceria com este ponto de equilíbrio à medida que afastamos o peso situado no ponto mais longe da origem para um local ainda mais afastado daquela? Não é difícil fazer a experiência, e ela pode, inclusive, ser simulada no computador; o resultado é que o ponto de equilibrio (a média) vai se afastando na direção daquele peso. Quanto mais afastamos o peso, o ponto de equilíbrio vai ficando cada vez mais longe do restante dos pesos que, desta forma, vão perdendo cada vez mais a importância para o cálculo da média. Imaginem se a distância deste peso mais afastado for um erro de medida. O valor da média ficaria totalmente errado.

Um outro exemplo deste ponto fraco da média: suponha que um dos seus tios ganhe o prêmio superacumulado da megassena. A renda média da sua família, contando a renda de cada um dos tios, passa a ser uma renda milionária, e qualquer colega que visse o resultado desta média acharia que você é muito rico. Na verdade, porém, o único rico é aquele tio que ganhou na megassena; você continua como antes.

Resumindo, a média é um ótimo resumo das grandezas, desde que não haja algum resultado que seja totalmente fora de proporção em relação à maioria.

5.2 Mediana

A mediana é o valor central dos dados. Da mesma forma que a média, a mediana somente existe para dados de variáveis quantitativas.

Para calculá-la, primeiramente temos de reorganizar os dados em ordem crescente e, em seguida, escolher o valor central. Se o número de dados for ímpar, então este valor central é único; se for par, fazemos a média dos dois valores centrais. A mediana pode também ser posta numa régua de uma forma análoga à que fizemos com a média, porém, neste caso, a mediana não vai equilibrar os pesos dos dois lados como a média, a não ser que as duas tenham o mesmo valor, o que somente acontece nos casos de simetria, mas isto deixamos para depois.

Podemos ver, pela forma como é calculada, que a única grandeza usada pela mediana de forma direta é o valor da medida central. As grandezas das outras variáveis são usadas somente de forma indireta, quando organizamos os dados em ordem crescente. Por isto, o que a mediana perde ao não usar todas as grandezas, ela ganha ao não sofrer do mesmo ponto fraco da média, ou seja, se um dos valores aumenta muito em relação aos outros, o valor da mediana permanece estável, não muda.

Usando os exemplos da média, se agrupamos nossos pesos na régua e marcamos a mediana e, da mesma forma, vamos arrastando um ponto para cada vez mais longe da origem, a mediana permanece imutável. Se um dos seus tios ganhasse na megassena, a renda média da sua família, calculada desta vez pela mediana, não mudaria, e vocês não virariam todos ricos de mentira, como no caso da média. Neste caso da renda familiar, a mediana é uma medida de tendência central mais representativa do que a média.

Um dos grandes problemas da mediana é a dificuldade de cálculo. Quando o número de dados é pequeno, não há problema. Agora imaginem ter de organizar, em ordem crescente, 100 números, 1000 números ou 1 milhão deles. Com os computadores atuais podemos fazer isto rapidamente até uns dez mil; porém 1 milhão fica difícil até para um computador. Por esta razão é que a média tem sido mais usada do que a mediana.

5.3 Média podada

Existem outras medidas, intermediárias entre a média e a mediana, chamadas de médias podadas. Elas não são muito usadas na estatística básica, porém ajudam na compreensão do conceito de média e de mediana.

Suponhamos, a princípio, que organizamos os dados em ordem crescente, da mesma forma que fizemos para calcular a mediana, porém, em vez de calcular a mediana, separamos, por exemplo, os 10% maiores e os 10% menores, e tiramos a média dos 80% dos valores que sobraram no meio. Este resultado é chamado de média podada a 10%. Se temos certeza de que no máximo 10%, ou menos, dos maiores e 10%, ou menos, dos menores, são resultados duvidosos (medidas malfeitas, por exemplo) esta média não vai sofrer do ponto fraco da "média"e vai usar plenamente 80% das grandezas dos dados. O nome desta medida será média podada a 10%.

Também vamos ter menos cálculos para fazer em relação à mediana pois, no fundo, não há necessidade de ordenar todos os dados, basta iniciar a ordenação por baixo e parar quando tiver ordenado os primeiros 10%, e, depois, fazer o mesmo com a ordenação por cima. Também é fácil ver que podemos ter médias podadas a 10%, ou 20%, ou 45%. Se examinamos ainda mais atentamente, vemos que a média como a conhecemos é a média podada a 0% e a mediana é a média podada 50%.

Em inglês, média podada se diz *trimmed mean*, assim, o comando para se fazer média podada no R é:

```
mean(x, trim = 0.15)  # 15% de poda
```

```
[1] 465.9
```

```
mean(x, trim = 0.4)  # 40% de poda
```

```
[1] 441.8
```

5.4 Moda

A moda é simplesmente o valor que mais vezes aparece no nosso conjunto de dados. Se todos os valores aparecem um número igual de vezes (em geral, uma vez cada), dizemos que nossos dados não têm moda, ou seja, a moda pode, muitas vezes, não existir. Uma outra particularidade importante da moda é que ela é a única, das três medidas de tendência central, que pode ser calculada quando a variável medida é qualitativa. Se olhamos para um gráfico de colunas, a moda é justamente a categoria que corresponde à coluna mais alta.

5.5 Cálculos da média, mediana e moda a partir de uma tabela de distribuição de frequência

Muitas vezes, conjuntos de dados, resultados de medidas de uma variável quantitativa contínua, são repassados para você já organizados numa tabela de distribuição de fequência. Isto quer dizer que você não terá acesso aos dados originais, mas somente às frequências dos intervalos de classe. Mesmo assim, como já foi dito em outro capítulo, se o número de intervalos é correto, perdemos muito poucas informações que os dados contêm. Neste caso, ao calcular as medidas de tendência central a partir da tabela de distribuição de frequências, podemos ter certeza de que os resultados serão aproximações boas e confiáveis das medidas calculadas a partir dos dados originais.

Para calcular estas medidas, primeiramente, já que não conhecemos os valores originais, assumimos que todos os valores dos dados pertencentes a um dado intervalo de classe são iguais ao ponto médio (pm) deste intervalo, o que quer dizer que cada valor é repetido tantas vezes quanto seja a frequência do intervalo ao qual ele pertence. Depois calculamos a média para estes valores. Fazemos o mesmo para a mediana.

Vamos usar a tabela 3.2, agregando o ponto médio e usando somente a frequência absoluta.

Tabela 5.1: Tabela de distribuição de frequências

Intervalo de classe	pm	f
[1, 2)	1.5	0
[2, 3)	2.5	1
[3, 4)	3.5	3
[4, 5)	4.5	20
[5, 6)	5.5	21
[6, 7)	6.5	5

Olhando a tabela, vemos que o valor 2.5 aparece uma vez, 3.5 aparece 3 vezes, 4.5 aparece 20 vezes e assim por diante. Usando a fórmula para o cálculo da média, temos:

$$\bar{x} = \frac{2.5+3.5+3.5+3.5+4.5+4.5+...+6.5}{1+3+20+21+5} = \frac{2.5x1+3.5x3+...+6.5x5}{50} = 5.02$$

Média

Com isso podemos deduzir a fórmula para a média a partir da tabela de distribuição de frequência.

Seja k o número de intervalos de classe e n o número total de dados. Seja também y_j para $j = 1 \ldots k$ os pontos médios dos intervalos de classe e f_j para $j = 1 \ldots k$ as frequências absolutas de cada intervalo. Então o valor de $\bar{X}$, usando somente a tabela de distribuição de frequência, é:

$$\frac{\sum_{j=1}^{k} y_j f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} y_j f_j}{n}$$

Código no R para se calcular a média a partir da tabela de frequência usando o pacote fdth para gerar a tabela de distribuição de frequência.

```
set.seed(3051952)
x ← rnorm(n = 50, mean = 5, sd = 2)
```

```
# gera 50 valores aleatórios a partir de uma variável normal
d ← fdt(x, start = 0.5, end = 8.5, h = 1)
d$table[, 1]  #primeira coluna da tabela (d$table é um data
    frame)
```

```
[1] [0.5,1.5) [1.5,2.5) [2.5,3.5) [3.5,4.5) [4.5,5.5) [5.5,6.5)
[7] [6.5,7.5) [7.5,8.5)
8 Levels: [0.5,1.5) [1.5,2.5) [2.5,3.5) [3.5,4.5) ... [7.5,8.5)
```

```
b ← apply(as.matrix(d$table[, 1]), 2, function(x) paste("mean(c
    (", substr(x,
    2, 9), ")"))
# substr retira o colchete e o paste cola a função mean(c(
b
```

```
     [,1]
[1,] "mean(c( 0.5,1.5) )"
[2,] "mean(c( 1.5,2.5) )"
[3,] "mean(c( 2.5,3.5) )"
[4,] "mean(c( 3.5,4.5) )"
[5,] "mean(c( 4.5,5.5) )"
[6,] "mean(c( 5.5,6.5) )"
[7,] "mean(c( 6.5,7.5) )"
[8,] "mean(c( 7.5,8.5) )"
```

```
a ← apply(b, 1, function(x) eval(parse(text = x)))
a
```

```
[1] 1 2 3 4 5 6 7 8
```

```
tab ← data.frame(Classlimits = d$table[, 1], pm = a, f = d$
    table[, 2])
```

1. "*mean*(c(0.5,1.5))"é uma expressão de caracteres.
2. *parse* cria um objeto tipo expressão.
3. *eval* calcula o resultado da expressão que, neste caso, vale 1 = média de 0.5 e 1.5.

4. A função *apply* aplica a cada linha de b a função especificada e calcula

5. O resultado do calculado é igual ao vetor *a*= 1 2 3 4 5 6 7 8 .

Finalmente, usando a função xtable do pacote do mesmo nome, fazemos a tabela com ponto médio.

```
library(xtable)
tx ← xtable(tab, caption = "Tabela com o ponto médio", label =
    "tab:pm")
print(tx, table.placement = "H", caption.placement = "top",
    latex.environments = "flushleft")
```

Tabela 5.2: Tabela com o ponto médio

	Classlimits	pm	f
1	[0.5,1.5)	1.00	2
2	[1.5,2.5)	2.00	2
3	[2.5,3.5)	3.00	5
4	[3.5,4.5)	4.00	11
5	[4.5,5.5)	5.00	7
6	[5.5,6.5)	6.00	7
7	[6.5,7.5)	7.00	7
8	[7.5,8.5)	8.00	6

Mediana

Da mesma forma como fizemos no caso da média, vamos considerar os pontos médios como se fossem os dados originais. Neste caso usamos a fórmula de mediana e encontramos o valor da mediana. Por exemplo, a tabela 5.1 foi gerada por uma amostra de 50 objetos (soma das frequências). Este número é par, logo a mediana é a média dos valores situados nas posições 25 e 26 dos dados ordenados. Os dois pertencem ao intervalo de classe [5, 6), logo a mediana é dada por $(5.5 + 5.5)/2 = 5.5$.

Moda

Para calcular a moda, primeiro encontramos o intervalo modal, que é o intervalo de maior frequência. O ponto médio deste intervalo é a moda. Preste atenção: a moda pode não ser única, pois uma tabela de distribuição de frequências pode ter mais de um intervalo modal. No caso da tabela 5.1, o intervalo [5, 6) é o intervalo modal, com $f = 21$, logo a moda é igual a 5.5.

5.6 Exercícios

1. Os dados gerados por:

```
sample(45:85, 120, replace = TRUE)
```

```
  [1] 52 75 83 55 80 57 57 84 79 79 77 47 52 81 78 60 52 80 51 73 82
 [22] 63 50 82 77 69 83 55 64 85 45 62 68 70 59 83 68 71 47 75 73 64
 [43] 59 74 46 85 71 79 60 70 49 80 66 49 62 49 47 72 73 62 62 73 70
 [64] 65 65 77 46 62 85 52 52 49 63 77 68 56 66 84 48 81 47 62 83 71
 [85] 78 71 63 45 70 52 74 64 54 73 84 55 51 46 63 82 75 56 79 61 84
[106] 59 59 77 75 60 68 84 45 77 49 63 66 52 85 46
```

representam os resultados da variável peso medida em uma amostra de 120 mulheres, coletada numa dada região do Brasil. Faça uma tabela de distribuição de frequências, usando o **fdth** com 7 intervalos de tamanho 6, começando a partir de 45. Calcule a média e a mediana.

depois faça uma tabela semelhante, porém com 10 intervalos de tamanho 4.

Capítulo 6

Medidas de dispersão ou variabilidade

Podemos dizer, sem sombra de dúvida, que o conceito de dispersão ou variabilidade é o conceito mais importante da Estatística, é a essência mesmo do pensamento estatístico.

Estatística é a ciência da organização da informação, no seu sentido mais amplo, sob a presença de incerteza, e a incerteza se manifesta, na prática, na forma de dispersão dos dados obtidos, e é usando esta variabilidade que damos credibilidade aos resultados estatísticos. Por exemplo, quando a variabilidade dos dados é alta, a precisão da informação que queremos alcançar é baixa. Em outras palavras, na presença de muita incerteza, temos um baixo nível de precisão nos nossos resultados, e vice-versa.

Com isso vemos que a variabilidade é a unidade de medida básica da Estatística. Da mesma forma que duas cidades estão longe uma da outra, se elas estão a muitos quilômetros uma da outra, dois resultados de medida de uma variável estarão longe se estão a muitas medidas de dispersão entre si. Voltaremos a este conceito mais adiante.

Falando com números: suponhamos que queremos fazer um estudo

sobre o peso e sobre a altura de jovens na idade entre 13 e 15 anos, e usamos uma classe de alunos de uma dada escola como amostra. Estamos supondo que esta classe de alunos é representativa da população a qual queremos estudar. Esta classe tem 30 alunos, quer e medimos o peso e a altura de cada um deles e estes são os nossos dados.

Qual será que tem uma maior variabilidade? Em outras palavras, qual das duas variáveis tem os resultados das medidas mais dispersos, mais heterogêneos, a da altura dos alunos ou a do peso?

- Se for a do peso, isto quer dizer que os alunos têm alturas com pouca variação e, neste caso, existem alunos magrinhos, menos magrinhos, gordinhos e mais gordinhos. Portanto, se olharmos por cima, veremos uma variação na altura das cabeças menor do que se olharmos de lado, vendo a largura dos corpos.

- Se for a altura, teremos alunos mais altos e magros e mais baixos e gordos para manter a menor dispersão nos pesos.

Por esta razão é de suma importância saber medir esta variabilidade. Se estudamos distância, temos de saber medir distâncias; se estudamos pressão atmoférica ou pressão arterial, temos de saber medir pressão; se estudamos corrente elétrica, temos de saber medir corrente elétrica; da mesma forma, se estudamos variabilidade ou dispersão, temos de saber medir variabilidade.

Como encontramos estas medidas de dispersão? Elas são calculadas a partir dos dados que coletamos. Aplicamos uma fórmula a estes dados e obtemos o resultado. Para que possam ser medidas de variabilidade, o resultado destas fórmulas tem de satisfazer a duas propriedades:

1. O resultado tem de ser sempre maior ou igual a zero, não importa quais sejam os números.

2. Se os dados são todos iguais, então o resultado da fórmula vale zero.

Isto porque não tem sentido falar em dispersão negativa, e se todos os dados são iguais, isto quer dizer que não existe dispersão, logo o resultado da medida tem de ser zero.

As duas formas básicas de calcular medidas de dispersão são a subtração de duas medidas de posição, das quais uma é sempre menor do que a outra. Por exemplo, o máximo menos o mínimo. A outra faz uma média de distâncias a uma medida de tendência central; a este tipo pertence a medida mais usada na estatística: o desvio padrão.

6.1 Medidas

Desvio padrão

O desvio padrão é calculado da seguinte forma: primeiro, calculam-se os desvios de cada dado da média, ou seja, o resultado da subtração do valor do dado menos o valor da média. Como os desvios podem ser positivos ou negativos, eleva-se ao quadrado cada um destes desvios para termos somente resultados positivos e somamos todos. Se somarmos somente os desvios da média, sem elevar ao quadrado, o resultado será sempre zero. Façam a conta com números e depois com letras, para comprovar e provar que o resultado zero é sempre válido. Finalmente, dividimos o resultado pela quantidade de dados menos um e temos o que chamamos de variância. Para encontrar o desvio padrão, basta tirar a raiz quadrada da variância. É fácil verificar que o desvio padrão tem a mesma unidade de medida dos dados: por exemplo, se os dados são em quilogramas ou em metros ou em dias, o desvio padrão terá a mesma unidade. A razão pela qual dividimos pela quantidade de dados menos um, e não somente pela quantidade de dados, faz parte da inferência estatística e não será discutida neste texto.

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}{n-1}}$$

```
set.seed(31051952)
```

```
x ← sample(1:1000, size = 20, replace = TRUE)

# no R a fórmula da variância é escrita assim:
sum((x − mean(x))^2)/(20 − 1)
```

```
[1] 69567
```

```
# que é igual a
(sum(x^2) − sum(x)^2/20)/(20 − 1)
```

```
[1] 69567
```

```
# esta segunda fórmula é mais fácil para computá-la numa
    calculadora
# simples. que é igual a
var(x)
```

```
[1] 69567
```

```
# o desvio padrão é igual a
sqrt(var(x))
```

```
[1] 263.8
```

```
# ou
sd(x)
```

```
[1] 263.8
```

Se os dados tiveram todos os mesmos valores, então a média terá exatamente este valor e todos os desvios serão zero, logo, também o desvio padrão. Em geral, como os quadrados dos desvios são todos positivos, assim também será o desvio padrão.

Separatrizes

Separatrizes são medidas calculadas a partir dos dados ordenados que servem para dividir o conjunto de dados em partes, a partir de

um percentual predefinido. No nosso caso, vamos dividir os dados em 4 partes, a partir de três valores chamados de quartis. Eles dividem o conjunto nos 25% menores, depois nos 25% entre aqueles e o centro, os 25% maiores que o centro e, finalmente, os 25% maiores. O centro, obviamente, é representado pela mediana, que também é o segundo quartil. Os quartis são calculados da seguinte forma:

1. Ordenamos o conjunto de dados.
2. Marcamos o mínimo, a mediana e o máximo.
3. Calculamos a mediana dos valores que são menores ou iguais à mediana. Este número é chamado de primeiro quartil.
4. Calculamos a mediana dos valores que são maiores ou iguais à mediana. Este número é chamado de terceiro quartil.

No R, para se calcular as separatrizes, e mais o mínimo, o máximo e a média de um conjunto de dados, usamos a função *summary* da seguinte forma:

```
set.seed(1121)
x ← c(sample(1:20, size = 15, replace = TRUE), 35)
summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0     8.5    12.0    13.8    18.0    35.0
```

Desvio interquartílico DIQ

Duas outras medidas de dispersão comumente usadas são: a diferença entre o máximo e o mínimo dos dados, chamada de amplitude; e a diferença entre o terceiro e o primeiro quartil, chamada de desvio interquartílico ou simplesmente DIQ.

Todas duas possuem as duas propriedades já citadas, como é fácil verificar. Se os dados são todos iguais, as medidas de posição também

serão. Em geral, o máximo é maior do que o terceiro quartil, que é maior que o primeiro quartil, que, por sua vez, é maior que o mínimo. Logo a amplitude e o desvio interquartílico são sempre maiores ou iguais a zero. É importante enfatizar que o DIQ tem a mesma unidade de medida dos dados da amostra.

Coeficiente de variação

O coeficiente de variação (CV) é uma medida de dispersão muito importante, pois é a única que permite comparar a variabilidade entre conjuntos de dados diferentes: no caso, medidas de variáveis quantitativas, quaisquer que sejam. Vamos explicar isto melhor. Suponhamos que queremos comparar a variabilidade de dois conjuntos de dados, a variável peso e a variável altura, as duas medidas na mesma amostra de objetos: neste caso, os alunos da classe de estatística básica. O valor do desvio padrão (dp) de cada uma vai depender da unidade de medida usada. Por exemplo, se o peso for em quilograma, o dp vale 8 kg, se o peso for em gramas o dp vale 8000g. Da mesma forma, o dp da altura vai variar de acordo com a unidade de medida, metros 0.20m ou centímetros 20cm. Fica clara a impossibilidade de saber qual das duas variáveis tem uma variabilidade maior, neste conjunto de dados, usando o dp como medida de comparação. Mesmo se os dados tiverem a mesma unidade de medida, mas não a mesma ordem de grandeza, não se pode comparar suas variabilidades usando o dp. Pense, por exemplo, no caso em que temos uma única variável, peso em gramas, sendo medida numa amostra de pessoas e em outra de formigas. Um dp de 0.3g é enorme para formigas, mas irrisório para seres humanos, ou seja, um mesmo dp mostra uma variabilidade enorme para formigas e uma praticamente nula para pessoas.

Isto mostra que é necessário uma medida de variabilidade que não tenha unidade de medida para comparar variabilidades de conjuntos de dados diferentes. Neste caso, definimos o coeficiente de variação como o desvio padrão dividido pela média e, depois, multiplicado por 100%.

$$CV = \frac{S}{\bar{X}} 100\%$$

```
# alt é um vetor de dados a partir da medida da variável altura
    em
# pessoas.
alt ← c(180, 181, 175, 183, 182, 165, 175, 171, 174, 180, 173,
    180, 183,
    187, 172, 183, 185, 175, 179, 190, 179, 170, 173, 168)

# hdl é um vetor de dados a partir da medida da variável
    colesterol
# em pessoas.
hdl ← c(37, 75, 35, 30, 62, 42, 43, 36, 51, 24, 41, 65, 31, 25,
     34, 30,
    37, 37, 43, 42, 33, 28, 49, 33, 58, 30, 40, 34, 38)

cv.alt ← sd(alt)/mean(alt) * 100
cv.alt
```

```
[1] 3.513
```

```
cv.hdl ← sd(hdl)/mean(hdl) * 100
cv.hdl
```

```
[1] 30.3
```

As duas variáveis têm unidades de medida diferentes; a altura é medida em centímetros, enquanto o colesterol é medido em miligramas por decilitro (mg/dL).

Usando o CV, podemos ver, no resultado acima, que o hdl tem uma variabilidade em relação à média, ou seja, em relação à ordem de grandeza dos seus valores, quase 9 vezes maior do que a de alt.

6.2 Distância estatística

Vamos agora estudar o conceito da distância estatística, ou seja, uma medida de distância que nos mostra quando dois valores estão estatisticamente longe ou perto um do outro.

Suponhamos que medimos a altura (variável) dos 30 alunos de uma certa classe. Calculamos a média e o desvio padrão que resultaram, respectivamente, 160 cm e 10 cm.

No caso da medida de distância comum, se dois objetos estão a meio metro de distância podemos dizer que estão perto um do outro. Da mesma forma, se duas medidas estão a meio desvio padrão uma da outra, podemos dizer que estão perto uma da outra, como, por exemplo, um aluno medindo 168 cm e outro medindo 173 cm estão estatisticamente perto um do outro. Da mesma forma, se dois alunos medem 162 e 182 cm estão estatisticamente longe um do outro. Em geral, uma grande parte dos dados está perto da média, ou seja, está a uma distância de um desvio padrão da média.

Vantagens e desvantagens

Da mesma forma que a média, o desvio padrão é muito sensível a valores muito maiores ou muito menores do que quase todos os dados. Por exemplo, suponhamos que os nossos dados são os números 1, 2, 3, 4, 5. O desvio padrão destes dados é 1,6. Se, por acaso, o último número, em vez de 5, é digitado 50, o desvio padrão passa a ser 21,3. Se o número passa a ser 500, o desvio padrão passa a ser 222,5. Se este número 500 apareceu devido a um erro de digitação, o desvio padrão será muito alterado.

Exatamente como no caso da média e da mediana, isto não acontece com o desvio interquartílico. Não importa o valor do quinto dado, ele será sempre igual a 4 - 2 = 2.

Por outro lado o desvio padrão é muito fácil de calcular e possui muito boas propriedades estatísticas, principalmente quando os dados

não contêm os valores estranhos citados acima.

6.3 Boxplot ou diagrama de caixa

O *boxplot* ou diagrama de caixa foi inventado pelo grande estatístico John Tukey, que foi o maior contribuidor para o desenvolvimento dos métodos e da metodologia chamada *Análise exploratória de dados*.

Diagrama de caixa

As separatrizes não são tão interessantes em si mesmas, contudo adquirem uma importância maior quando usadas para formular o diagrama de caixa.

```
par(mfrow = c(1, 2), cex = 0.75)
boxplot(x, main = "Variável x")
boxplot(mpg ~ cyl, data = mtcars, main = "Milhas por galão",
    xlab = "Número de cilindros",
    ylab = "Milhas por galão")
```

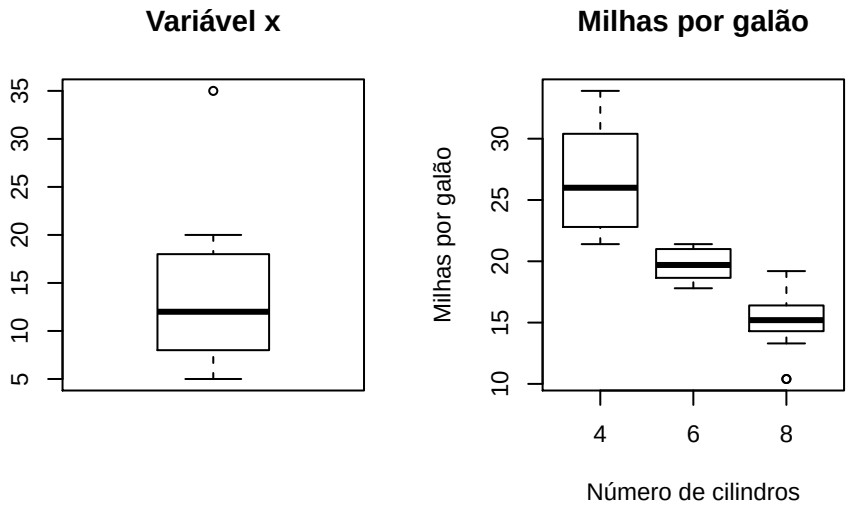


Figura 6.1: Diagramas de caixa

A partir da figura 6.1, podemos ver como se constrói o diagrama de caixa. Primeiramente, marcamos o lugar da mediana ou segundo quartil (2q) no eixo vertical. No nosso caso, do vetor x, ela vale 12 e está marcada com um segmento de reta horizontal em negrito. Depois, marca-se um outro segmento, correspondente ao primeiro quartil (1q), que vale 9. Da mesma forma, marcamos o segmento correspondente ao terceiro quartil (3q), igual a 18, e fechamos a caixa. A altura desta caixa, ou seja, 3q - 1q, no nosso caso $18 - 9 = 9$, é o desvio interquartílico. Finalmente marcamos o valor 20 com um segmento pequeno e, igualmente, o mínimo 6, e os ligamos à caixa com uma linha tracejada. O valor máximo é igual a 35, porém, como este valor está muito longe da maioria dos dados, o R o marca como valor atípico, e escolhe o valor anterior,20, para fechar o diagrama de caixa.

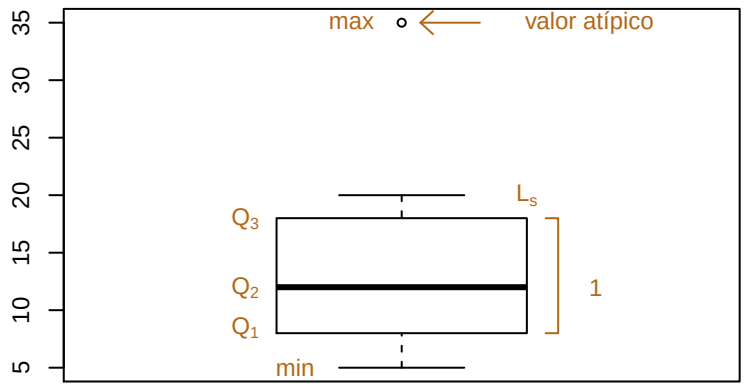


Figura 6.2: Diagramas de caixa explicado

No diagrama de caixa temos a ideia da variabilidade pelo DIQ, do valor central pela mediana e onde ela se situa em relação aos quartis, no nosso caso, ela está mais perto de 1q. Quando o máximo ou o mínimo estão muito longe dos respectivos quartis, em relação ao DIC, podemos supor que eles e alguns outros valores que estejam perto deles, sejam valores discrepantes ou atípicos da maioria dos dados. Em inglês, são conhecidos como *outliers*.

O diagrama de caixa também pode ser usado quando queremos comparar resultados para dados que podem ser comparados. Isto quer dizer que os dados têm a mesma unidade de medida e a mesma ordem de grandeza. No caso dos 3 diagramas de caixa da figura 6.1, temos a comparação de gasto de combustível, no caso milhas por galão de gasolina, para 3 diferentes números de cilindros, e podemos ver claramente que, quanto maior o número de cilindros, menor é a milhagem por galão.

Exercícios

1. Gere dados simulados que representem duas variáveis, por exemplo, peso e altura, e calcule todas as medidas de dispersão estuda-

das para cada um deles. Compare as duas variabilidades, usando a medida cabível.

2. Usando o *dataframe* de dados *mtcars*, faça o *boxplot* da variável (coluna) *hp*. Use a coluna *carb* para diferenciar os vários grupos da variável *hp* e faça os diagramas de caixa de cada uma. Compare todos os diagramas, inclusive o do *hp total*, seguindo o exemplo da figura 6.1.

Capítulo 7

Correlação e diagrama de dispersão

Frequentemente temos de analisar dados obtidos de duas variáveis que têm uma certa relação entre si. Um dos casos mais comuns de duas destas variáveis são o peso e a altura de pessoas, e as perguntas que nos fazemos são:

- Será que realmente existe uma relação, ou melhor dito, uma correlação entre o peso e a altura das pessoas?
- Se existe, qual é o seu significado estatístico e como isto poderá ser medido?

Vamos estudar somente o que chamamos de correlação linear entre duas variáveis. Duas variáveis podem ter outros tipos de correlação entre elas, como, por exemplo: quadrática, exponencial, logarítmica e muitas outras. Porém a mais fácil de estudar e conseguir bons resultados estatísticos é a linear, a que mais acontece na natureza, ou, pelo menos, em relações que podem ser aproximadas pela linear ou que, por meio de uma transformação das variáveis, podem ser linearizadas.

Primeiro vamos estudar a forma gráfica de ver duas variáveis e como inferir, a partir deste gráfico, o efeito da correlação entre elas. Antes de tudo, porém, temos de saber como obter a amostra de duas variáveis relacionadas entre si.

7.1 Amostragem de dados bivariados

Se medimos a variável altura num grupo de pessoas e a variável peso em outro grupo, estas duas medições são independentes uma da outra, portanto, assim também serão as amostras, ou seja, elas não trazem em si nenhuma informação sobre a correlação que porventura possa existir entre as duas variáveis citadas. Isto quer dizer que, se queremos obter informação sobre a correlação entre a duas variáveis, não podemos amostrar da forma descrita acima.

A forma de fazê-lo é a seguinte:

1. Coletamos a amostra dos objetos nos quais vamos medir as variáveis. Em cada um destes objetos,

    a) medimos a primeira variável.

    b) medimos a segunda variável.

Cada objeto guarda, em si, a relação entre as duas variáveis, logo os dois grupos de dados, além de guardar informações sobre cada variável em separado, também guarda informações sobre a correlação entre elas.

A forma de simular este tipo de dados no R é mais sutil do que simplesmente gerar dois conjuntos de dados, como tem sido feito até agora. A descrição e o código estão exemplificados a seguir.

O primeiro conjunto de dados (x,y) foi feito da seguinte forma: primeiro selecionamos os valores de x usando o *sample*, depois fizemos uma função linear de x, neste caso $y = 3x$ seguido de uma variabilidade. Usamos, para isto, a função que gera números aleatórios, seguindo a

distribuição normal, a mais apropriada para simular relações lineares entre variáveis.

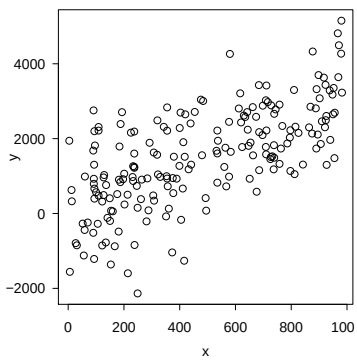
O segundo conjunto de dados simula duas variáveis (x1,y1); neste caso usamos $y = x^2$ seguido de uma variabilidade, mostrando uma relação quadrática entre as variáveis.

O primeiro diagrama simula a correlação linear, enquanto o segundo simula a correlação quadrática. Neste livro, iremos nos ater somente à correlação linear, que é a mais usada no mundo dos usuários da estatística.

```
set.seed(31051952)
x ← sample(1:1000, size = 200, replace = TRUE)
y ← 3 * x + rnorm(200, mean = 0, sd = 1000)
x1 ← sample(seq(1, 20, by = 0.01), size = 200, replace = TRUE)
y1 ← x1^2 + rnorm(200, mean = 0, sd = 15)
```

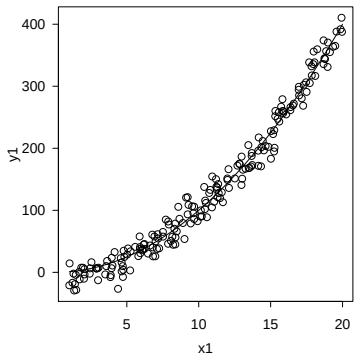
7.2 Diagrama de dispersão

```
par(mar = c(4, 4, 0.1, 0.1), cex.lab = 0.95, cex.axis = 0.9, mgp
    = c(2,
    0.7, 0), tcl = −0.3, las = 1)
plot(x, y)
```




```
plot(x1, y1)
```

```
v ← seq(1, 20, by = 0.01)
lines(v, v^2)
```



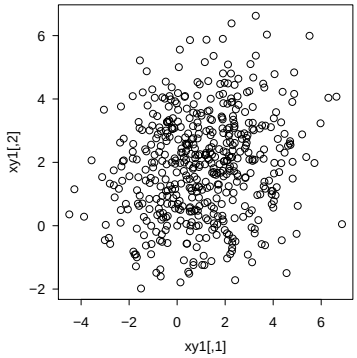
Os dois gráficos apresentados são chamados de **diagramas de dispersão**. No eixo das abscissas, eixo horizontal, marcamos os valores da variável x e o seu correspondente da variável y no eixo das ordenadas, ou eixo vertical. Desta forma, cada objeto corresponde a um ponto do diagrama bidimensional, cujas coordenadas são os valores das duas variáveis medidas neste objeto.

Outra forma de simular dados originados das medidas de duas variáveis com uma relação linear entre elas é usando o pacote mvtnorm, (GENZ et al., 2013), porém, para poder entendê-lo com mais profundidade, você deverá ter algum conhecimento da teoria das probabilidades.

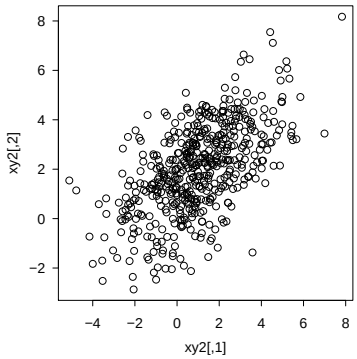
```
library(mvtnorm)
sigma1 ← matrix(c(4, 0.5, 0.5, 3), ncol = 2)
sigma2 ← matrix(c(4, 2, 2, 3), ncol = 2)
sigma3 ← matrix(c(4, 3.4, 3.4, 3), ncol = 2)
xy1 ← rmvnorm(n = 500, mean = c(1, 2), sigma = sigma1)
xy2 ← rmvnorm(n = 500, mean = c(1, 2), sigma = sigma2)
xy3 ← rmvnorm(n = 500, mean = c(1, 2), sigma = sigma3)
```

```
par(mar = c(4, 4, 0.1, 0.1), cex.lab = 0.95, cex.axis = 0.9, mgp
    = c(2,
```

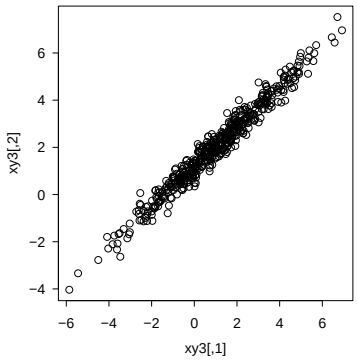
```
    0.7, 0), tcl = -0.3, las = 1)
plot(xy1)
```




```
plot(xy2)
```




```
plot(xy3)
```



Nestes três gráficos anteriores, vemos que o primeiro tem o formato mais arredondado; o seguinte, mais alongado; enquanto o terceiro é mais parecido com um charuto, com os pontos emaranhados ao redor de uma reta. No primeiro caso, dizemos que a correlação entre as duas variáveis é frouxa, ou quase inexistente; no segundo caso, ela é média, ou seja, as variáveis estão relacionadas linearmente, porém a correlação não é muito forte; no terceiro caso, podemos dizer que elas estão relacionadas linearmente e que a correlação é forte.

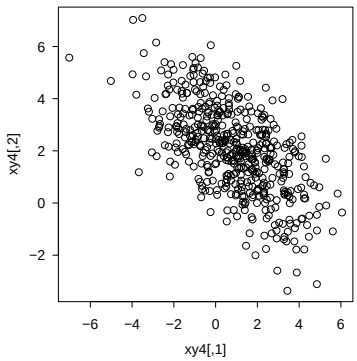
Em qual sentido, porém, definimos esta força na relação entre elas? Quando a correlação entre as variáveis é muito forte, podemos dizer que, dado o valor de x, poderíamos encontrar o valor de y a partir de uma equação linear $y = a + by$ e, ao compararmos com o valor real, encontrado ao medir a variável y, no mesmo objeto, erraríamos por muito pouco. À medida que a força desta relação se torna mais fraca, este erro aumenta, até se tornar tão grande a ponto de ofuscar completamente o valor de y, o que seria o caso do primeiro gráfico.

Por outro lado, a correlação linear pode ser tanto positiva como negativa. Se for positiva, a inclinação geral dos pontos tem a mesma inclinação de uma reta, cujo coeficiente angular é positivo. Se for negativa, a inclinação geral dos pontos tem a inclinação de uma reta com coeficiente angular negativo.

```
sigma4 ← matrix(c(4, −2, −2, 3), ncol = 2)
```

```
xy4 ← rmvnorm(n = 500, mean = c(1, 2), sigma = sigma4)
```

```
par(mar = c(4, 4, 0.1, 0.1), cex.lab = 0.95, cex.axis = 0.9, mgp
    = c(2,
    0.7, 0), tcl = -0.3, las = 1)
plot(xy4)
```



No caso da correlação positiva, entendemos que, à medida que os valores de uma variável crescem, os da outra também tendem a crescer. Isto não quer dizer que se $x_1 < x_2$ então $y_1 < y_2$. Se a força da relação linear for muito forte, as duas desigualdades acontecerão quase sempre, porém, se a força da relação for fraca, aquela formulação acontecerá mais raramente. Analogamente ocorre o mesmo quando a correlação for negativa, porém no sentido contrário, isto é: à medida que os valores de uma variável crescem, os da outra também tendem a diminuir.

7.3 Coeficiente de correlação

O coeficiente de correlação é um número que varia entre -1 e 1 e que nos informa tanto sobre a força como sobre o sinal da correlação linear entre duas variáveis. Este coeficiente foi idealizado por Karl Pearson no início do século 20 , por isto leva o nome de *coeficiente de correlação de Pearson*

Sejam X e Y as duas variáveis em questão, por exemplo, peso e altura. Sejam x_i $i = 1 \ldots n$ e y_i $i = 1 \ldots n$ os valores das variáveis medidas em cada objeto da amostra, ou seja, para cada pessoa pertencente à amostra medimos seu peso e sua altura correspondente. Então, definimos a covariância entre X e Y por:

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

A covariância é a medida que realmente mede a variação linear de uma variável em relação à outra. O problema com a covariância é que ela tem unidade de medida, que é o produto das unidades de medida de cada variável, e, por esta razão, ela pode ter diferentes grandezas dependendo das variáveis. Portanto, é muito mais útil usar uma medida que não tenha unidade, e seu valor seja estandartizado, isto é, que tenha os mesmos limites inferiores e superiores para quaisquer que sejam as variáveis. Para isto, utilizamos o coeficiente de correlação.

Para estandartizar a covariância, dividimos o seu valor pelo produto dos desvios padrões das duas variáveis. Desta forma definimos o *coeficiente de correlação*

$$r = \frac{cov(X,Y)}{S_x S_y}$$

No R, o cálculo da covariância e do coeficiente de correlação é feito da seguinte forma:

```
set.seed(31051952)
n ← 200
x ← sample(1:1000, size = n, replace = TRUE)
y ← 3 * x + rnorm(n, mean = 0, sd = 1000)

cov ← (sum((x − mean(x)) * (y − mean(y))))/(n − 1)
cov
```

```
[1] 263661
```

```
cov ← (sum(x * y) − n * mean(x) * mean(y))/(n − 1)
cov
```

```
[1] 263661
```

```
cor ← cov/(sd(x) * sd(y))
cor
```

```
[1] 0.6709
```

```
cov(x, y)
```

```
[1] 263661
```

```
cor(x, y)
```

```
[1] 0.6709
```

```
# Do resultado anterior, usando a função rmvnorm
cov(xy1[, 1], xy1[, 2])
```

```
[1] 0.5593
```

```
cor(xy1[, 1], xy1[, 2])
```

```
[1] 0.1762
```

O coeficiente de correlação, por ser estandartizado, torna-se fácil de ser interpretado. Não é difícil provar que o seu valor varia de -1 a 1 e, portanto, se for positivo, a correlação linear entre as duas variáveis é positiva; ou analogamente negativa. Por outro lado, quanto mais perto de zero, mais fraca é a relação entre as duas variáveis, e quanto mais perto de 1 ou -1, mais forte a relação linear entre elas.

7.4 Exercícios

1. Use a equação de uma reta qualquer para simular duas variáveis correlacionadas. Varie o desvio padrão na função *rnorm*, calcule os coeficientes de correlação e faça os diagramas de dispersão. Compare os resultados para ver a relação entre o diagrama de dispersão e o coeficiente de correlação.

2. No exemplo em que o pacote *mvtnorm* é usado, a matrix sigma1 é a matrix de covariância, onde o valor da variância de X vale 4; o valor da variância de Y vale 3 e o valor da covariância entre X e Y vale 0.5. O vetor *mean* contém as médias populacionais de X e Y.

   Se você souber usar a função *rmvnorm*, use-a, variando os valores das médias e da matriz sigma. Faça os diagramas de dispersão e calcule os coeficientes de correlação.

   Uma pergunta à parte: por que, se você calcular a média, a variância e a covariancia das amostras simuladas, o valor nunca será exatamente o mesmo dos valores usados na função *rmvnorm*? Serão parecidos mas não os mesmos.

Parte II

Dados multivariados

Capítulo 8

Análise de correspondência

Análise de correspondência (AC) é uma metodologia estatística voltada para a análise exploratória de dados categóricos multivariados. Veja, por exemplo, o livro (GREENACRE, 2007) que descreve o modelo de uma forma muito didática.

Existem duas formas de exibir dados categóricos multivariados para serem usados na AC. A primeira usando uma tabela de contingência quando utilizamos somente duas variáveis, e a segunda, na forma de uma matriz, ou na linguagem do R um *data.frame*, na qual cada linha corresponde a um sujeito (ou objeto) da amostra, e cada coluna corresponde a uma variável.

No caso da primeira forma, usamos a AC simples e, no caso da segunda, usamos a AC múltipla. Como foi comentado anteriormente, a AC nos ajuda a compreender as relações e associações existentes entre as variáveis. Em resumo, o que a AC faz é reduzir a dimensionalidade do espaço destas variáveis, projetando-as num gráfico de duas dimensões. Seu resultado é uma representação gráfica, simples e elegante, que leva a uma rápida interpretação e ao entendimento da estrutura por trás dos dados. Em outras palavras, a análise de correspondência simplifica a complexidade de uma alta dimensionalidade, descrevendo toda a informação contida nos dados. Quando usamos dados multivariados, cada

variável pode ser representada como uma dimensão, logo, se estamos tratando de sete variáveis, estaremos trabalhando num espaço de sete dimensões, e isto torna a análise extremamente complexa. A AC usa um método de álgebra linear, chamado decomposição em valores singulares, para mudar as coordenadas do espaço usual de várias dimensões para outras que têm a direção de maior variabilidade, depois a segunda de maior variabilidade, e assim por diante, em ordem decrescente. Neste caso, se projetamos os pontos relativos a cada variável no espaço de duas dimensões, gerado pelas duas primeiras coordenadas descritas, isto é, as duas na direções de maior variabilidade, esta projeção irá conter uma grande parte de toda a informação contida nos dados. Não se preocupe se você não conseguiu entender esta descrição. Ao estudar os exemplos, você irá entender do que se trata. Para fazer todos os cálculos e gerar as tabelas e os gráficos necessários, e assim obter os resultados necessários, vamos usar o pacote do R chamado *ca* (NENADIC; GREENACRE, 2007).

O conjunto de dados abaixo, chamado *smoke*, contém frequências de hábitos de fumar (nenhum, pouco, moderado e forte) para a equipe de administração (gerente sênior, gerente júnior, funcionário sênior, funcionário júnior e secretárias) numa companhia. Os dados são fictícios.

```
library(ca)
library(xtable)
data(smoke)
# Dando nomes às linhas e colunas do dataframe smoke.
colnames(smoke) <- c("nenhum", "pouco", "moderado", "forte")
rownames(smoke) <- c("GS", "GJ", "FS", "FJ", "SC")
```

```
# Usando o pacote xtable para construir a tabela em Latex
print(xtable(tab, caption = "Hábitos de fumo", label = "tab:
    corresp2"),
    table.placement = "H", latex.environments = "flushleft",
        hline.after = c(-1,
        0))
```

	Classlimits	pm	f
1	[0.5,1.5)	1.00	2
2	[1.5,2.5)	2.00	2
3	[2.5,3.5)	3.00	5
4	[3.5,4.5)	4.00	11
5	[4.5,5.5)	5.00	7
6	[5.5,6.5)	6.00	7
7	[6.5,7.5)	7.00	7
8	[7.5,8.5)	8.00	6

Tabela 8.1: Hábitos de fumo

```
# Usando o pacote ca para fazer a análise de correspondência
smoke.ca ← ca(smoke)
summary(smoke.ca)
```

```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.074759  87.8  87.8  ************************
 2      0.010017  11.8  99.5  ***
 3      0.000414   0.5 100.0
        -------- -----
 Total: 0.085190 100.0


Rows:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |   GS |   57  893   31 |  -66  92   3 | -194 800 214 |
2 |   GJ |   93  991  139 |  259 526  84 | -243 465 551 |
3 |   FS |  264 1000  450 | -381 999 512 |  -11   1   3 |
4 |   FJ |  456 1000  308 |  233 942 331 |   58  58 152 |
5 |   SC |  130  999   71 | -201 865  70 |   79 133  81 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
```

```
1 | nnhm |  316 1000  577 | -393 994 654 |  -30   6  29 |
2 | pouc |  233  984   83 |   99 327  31 |  141 657 463 |
3 | mdrd |  321  983  148 |  196 982 166 |    7   1   2 |
4 | fort |  130  995  192 |  294 684 150 | -198 310 506 |
```

```
par(cex = 0.7, mar = c(2.5, 3, 2, 0.8) + 0.1)
plot(smoke.ca)
```

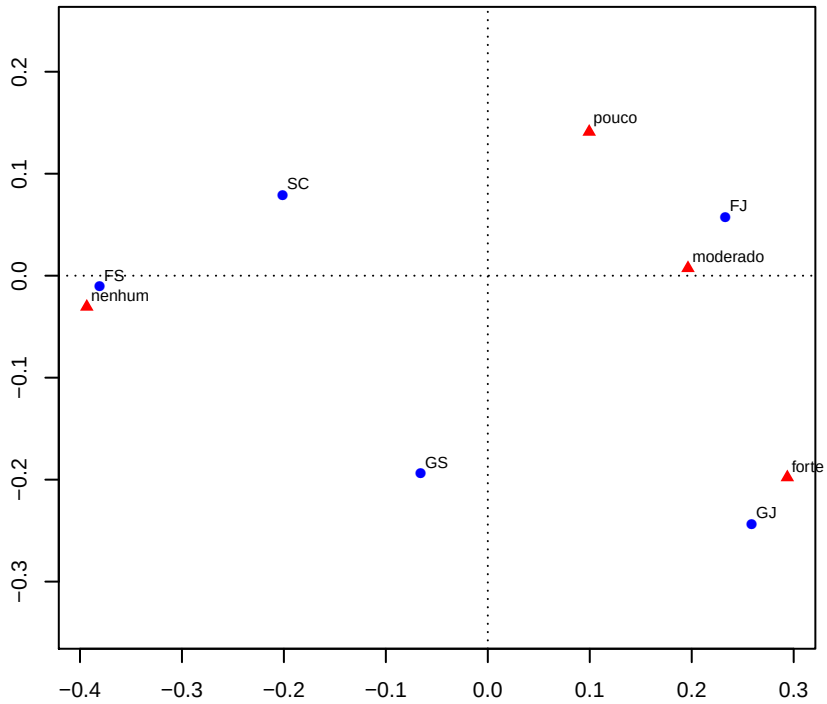


Figura 8.1: Resultado gráfico da AC

```
# Proporção de cada linha da tabela 1 do summary em relação ao
    seu
```

```
# total, ou seja, os valores de cada linha estão divididos pelo
# respectivo total.
round(t(apply(smoke, 1, function(x) x/sum(x))), 2)
```

```
   nenhum pouco moderado forte
GS   0.36  0.18     0.27  0.18
GJ   0.22  0.17     0.39  0.22
FS   0.49  0.20     0.24  0.08
FJ   0.20  0.27     0.38  0.15
SC   0.40  0.24     0.28  0.08
```

```
# Proporção de cada coluna da tabela 2 do summary em relação ao
    seu
# total, ou seja, os valores de cada coluna estão divididos pelo
# respectivo total.
round(apply(smoke, 2, function(x) x/sum(x)), 2)
```

```
   nenhum pouco moderado forte
GS   0.07  0.04     0.05  0.08
GJ   0.07  0.07     0.11  0.16
FS   0.41  0.22     0.19  0.16
FJ   0.30  0.53     0.53  0.52
SC   0.16  0.13     0.11  0.08
```

Vamos, então, analisar o resultado da AC. Primeiramente, vamos olhar o resultado do "summary".

Pelo fato de a tabela de contingência dos dados ter 5 linhas e 4 colunas, a dimensão gráfica máxima da AC é igual a 3, o menor valor entre número de linhas e número de colunas, menos um. No nosso caso, temos esta dimensão máxima igual a 3, que é também o número de inércias principais. Estas inércias principais dão o percentual da informação total explicada por cada dimensão. Como os gráficos são sempre de duas dimensões, estamos interessados nas duas primeiras inércias principais. No nosso caso, as duas primeiras inércias explicam 99.5% de toda a informação contida nos dados, o que representa quase tudo. A primeira delas explica 87.8%, e a segunda, 11.8%.

A segunda parte do "summary" mostra alguns resultados para as linhas (*rows*) e os mesmos resultados para as colunas (*columns*). A massa (*mass*) representa a proporção do total de cada linha (coluna) em relação ao total geral multiplicado por mil, ou , $\frac{11}{193}1000 = 57$, sendo 11 o total da linha GS, e 193 o total geral. A "qlt"representa a qualidade da representação de cada linha (coluna) de um total de 1000. Isto significa a proporção da informação total, que está contida em 3 dimensões, explicada por cada linha (coluna) do gráfico de duas dimensões. Estes números são muito altos, visto que a inércia total explicada pelas duas dimensões também é muito alta. A "inr"representa a inércia de cada linha (coluna) como proporção do total geral. A soma das inércias das linhas (colunas) é igual a 1000. Esta "inr", junto com a massa, dá uma ideia da importância de cada linha (coluna) na construção da AC.

Em geral, quando linhas (colunas) têm "mass"muito pequena e "inr" muito grande, estas linhas (colunas) são consideradas valores discrepantes (*outliers*) e deveriam ser descartadas da AC.

Agora vamos analisar o resultado gráfico da AC, na figura 8.1.

A componente principal mais importante, a que explica a maior parte da inércia, corresponde ao eixo horizontal e a segunda, ao vertical. Em relação à quantidade de cigarros, vemos que "pouco", "moderado"e "forte"estão de um lado do eixo horizontal, enquanto "nenhum"se encontra do outro lado, em relação à origem. Isto quer dizer que "nenhum"é bem diferente na sua composição dos outros três. As principais diferenças ocorrem entre "nenhum"e "forte", já que, projetados sobre o eixo horizontal, estão mais longe da origem. Quanto mais perto da origem está um ponto, mais ele se parece com as proporções médias. As proporções médias são os valores totais divididos pelo total geral. Usando a Matemática, isto quer dizer que o vetor de "pouco"**(2 3 10 24 6)** dividido por 45 é mais parecido com o vetor dos totais **(11 18 51 88 25)**, dividido por 193, do que o vetor de "nenhum" **(4 4 25 18 10)**, dividido por 61. Se todos os pontos estivessem na origem, então teríamos o caso de as duas variáveis serem independentes.

Em relação à equipe de administração, vemos que FS e SC estão

mais para a esquerda, enquanto que FJ e GJ estão para a direita no eixo horizontal, o que mostra que as principais diferenças estão entre eles.

Em relação às duas variáveis, podemos ver que FS não fumam nenhum cigarro em sua maioria. Pode-se ver, no resultado "**proporção de cada linha da tabela 1 do summary em relação ao seu total**", que a proporção de FS nas linhas (0.49) é a maior de todas na linha FS; por outro lado, a proporção de FS nas colunas da tabela, em relação ao total de cada coluna, também tem a maior proporção de FS na coluna "nenhum", (0.41). Os FJ estão entre fumantes moderados e poucos, e pode-se ver, naqueles mesmos resultados, que na linha FJ as maiores proporções estão em "moderado"(0.38) e "pouco"0.27; em relação às colunas, a proporção de FJ é a mesma nas colunas "pouco"e "moderado"; o FJ está mais perto do "moderado"por causa da diferença nas primeiras proporções. Os dois juntos somam bem mais da metade do total de FJ. Os GJ são fumantes fortes, neste caso, a linha GJ tem as proporções de 0.39 em "moderado"e 0.22 em "forte". No gráfico, o ponto GJ ficou mais perto do "forte", pelo fato de que se olharmos as colunas "moderado"e "forte"a proporção de GJ é maior em "forte"(0.16) do que em "moderado"(0.11).

É importante lembrar que toda a análise feita nas tabelas de dados não poderia ter sido feita sem que antes se tivesse analisado o gráfico da AC. Imaginem então quando a tabela de contingência for muito maior, ou seja, com muitas colunas e linhas, seria praticamente impossível analisá-la com detalhes somente inspecionando a tabela.

8.1 Análise de correspondência múltipla, ACM

A ACM é usada no caso de querermos analisar várias variáveis no lugar de duas como na AC. Neste caso, o formato da tabela dos dados é o seguinte:

- As variáveis são listadas nas colunas da tabela.
- Cada linha corresponde à resposta de cada objeto, ou sujeito da amostra, a todas as variáveis.

- As respostas marcam os níveis de cada variável.

Vejam o exemplo a seguir.

Neste exemplo, o nosso interesse principal está nas colunas onde estão as variáveis e não as linhas que correspondem aos objetos da amostra.

Os dados correspondem a uma tabela dos resultados de 488 estudantes de uma universidade. Cada estudante foi medido segundo as variáveis concorrrência no vestibular (conc), área de conhecimento (area), taxa de aprovação (tapr), média geral (craa). Os níveis de cada variável são:

- conc1=5 a 10; conc2=10 a 15; conc3=15 a 20; conc4=maior que 20
- area1=exatas; area2=humanas; area3=vida
- tapr1=menor que 50%; tapr2=entre 50% e 75%; tapr3=maior que 75
- craa1=menor que 5; craa2=entre 5 e 7,5; craa3=maior que 7,5

Ao olharmos o gráfico resultante da ACM, o que mais chama a atenção é que, ao projetarmos todos os pontos no eixo horizontal, o mais importante dos dois, vemos que todas as variáveis que têm uma certa ordem (conc, tapr, area) caminham do menor para o maior nível, da esquerda para a direita, o que mostra que o que influencia a relação entre as variáveis é a qualidade dos estudantes quanto ao resultado, ou seja, a variação vai dos maus alunos até os bons alunos. Melhor dizendo, os maus alunos (tapr1, craa1, tapr2, conc1) estão relacionados com a área de exatas, tudo isto à esquerda da origem. No meio, estão os alunos medianos (craa2, conc2), que estão relacionados com a área de humanas. À direita da origem, estão os bons alunos (conc3, tapr3, craa3, conc4), que estão relacionados à área da vida. Poderíamos, então, discutir sobre o porquê de, nesta universidade, os alunos de exatas ficarem como maus alunos, enquanto os da vida como bons alunos, e os de humanas como alunos medianos. Isto, porém, envolve um conhecimento maior dos cursos, dos alunos e da estrutura da universidade.

```
# Usando função mjca do pacote ca para fazer a análise de
# correspondência mltipla
cpv ← read.table("C:/LivroExploratoria/cpv2.txt", header = TRUE
    )
cpv.ca ← mjca(cpv)
# summary(cpv.ca)
```

```
par(cex = 0.7, mar = c(2.5, 3, 2, 1) + 0.1)
plot(cpv.ca)
```

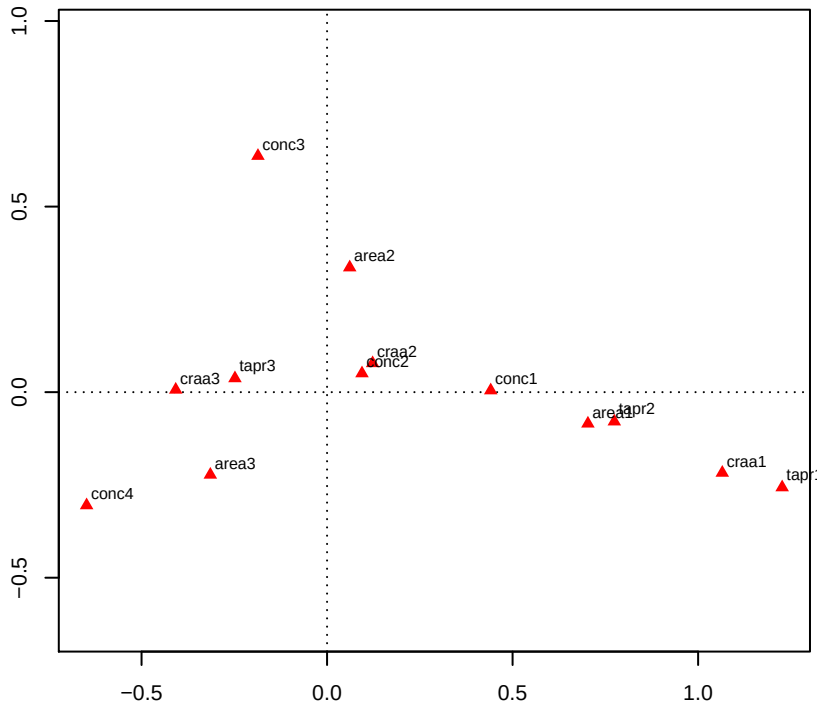


Figura 8.2: Resultado gráfico da ACM

Capítulo 9

Biplots

Biplot é uma metodologia estatística voltada para a análise exploratória de dados quantitativos multivariados.

Da mesma forma que a AC, o *biplot* também reduz a dimensionalidade dos dados para um gráfico em duas dimensões. Os dados são apresentados no formato de uma tabela, no R um *data.frame*, na qual qual as colunas representam as variáveis e as linhas representam os objetos da amostra, ou seja, os dados, em cada linha, são as medidas das variáveis no objeto correspondente àquela linha, ou seja, as observações. Estes *data.frames* são o material resultante de várias áreas de pesquisa. As linhas são indivíduos, países, grupos demográficos, lugares, casos, e as colunas são as variáveis que descrevem as linhas, como respostas de questionários, indicadores econômicos, produtos comprados, parâmetros ambientais, marcadores genéticos etc. Para uma descrição muito boa do método, veja o livro (GREENACRE, 2010).

A ideia básica do *biplot* é simples e, como todas as soluções simples de problemas complexos, é ao mesmo tempo poderoso e muito útil. O *biplot* faz com que a informação de uma tabela de dados se torne transparente, revelando as principais estruturas dos dados, de uma forma metódica como, por exemplo, padrões de correlação entre variáveis e similaridades entre observações. O pacote do R usado para obtermos os resultados

necessários é o bpca (FARIA; DEMETRIO, 2013).

9.1 Doze países da Europa

```
# Chamando os pacotes bpca e xtable
library(bpca)
library(xtable)
# Lendo os dados dos 12 países da Europa para um data.frame
eur ← read.table("C:/LivroExploratoria/livro/dataEuropa.txt",
    header = TRUE)
```

```
# Usando o pacote xtable para construir uma tabela em Latex a
    partir
# do data.frame eur
print(xtable(eur, caption = "12 países da Europa", label = "tab:
    biplot2"),
    caption.placement = "top", latex.environments = "flushleft",
        hline.after = c(-1,
        0), add.to.row = list(pos = list(12), command = c("\\
            hline\\multicolumn{4}{p{6cm}}\n{\\footnotesize\nX1=
            poder de compra per capita em euros.\nX2= Produto
            interno bruto(PIB) per capita.\nX3= Taxa de inflação
             (percentual)}")))
```

```
# Usando o pacote bpca para construir o gráfico de biplot
rownames(eur) ← eur$Pa.abrv.
par(mar = c(4, 4, 2, 1) + 0.1)
plot(bpca(eur[, 3:5]), obj.cex = 0.5, var.cex = 0.5)
```

Tabela 9.1: 12 países da Europa

	Pa.abrv.	Paises	X1	X2	X3
1	Be	Bélgica	19200	115.20	4.50
2	De	Dinamarca	20400	120.10	3.60
3	Ge	Alemanha	19500	115.60	2.80
4	Gr	Grécia	18800	94.30	4.20
5	Sp	Espanha	17600	102.60	4.10
6	Fr	França	19600	108.00	3.20
7	Ir	Irlanda	20800	135.40	3.10
8	It	Itália	18200	101.80	3.50
9	Lu	Luxemburgo	28800	276.40	4.10
10	Ne	Holanda	20400	134.00	2.20
11	Po	Portugal	15000	76.00	2.70
12	UK	Inglaterra	22600	116.20	3.60

X1= poder de compra per capita em euros.
X2= Produto interno bruto(PIB) per capita.
X3= Taxa de inflação (percentual)

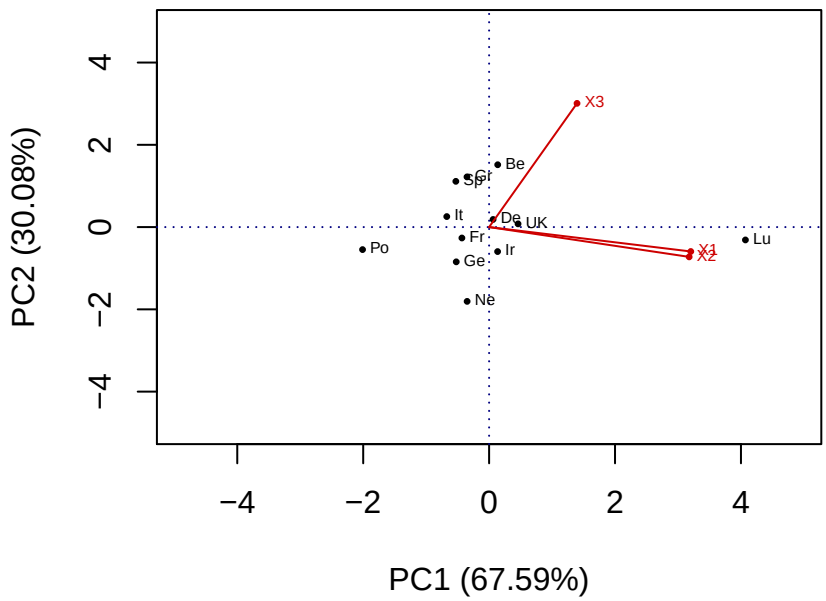


Figura 9.1: Biplot, resultado gráfico

A figura 9.1 mostra o *biplot* da tabela 9.1; foi usado o pacote bpca do R. Podemos ver que esta representação, somente em duas dimensões, consegue explicar 97,7% (67,6% + 30,1%) da informação total dos dados. Nesta tabela, os países são os objetos, e X1, X2 e X3 as variáveis. Os pontos que representam as variáveis estão ligados à origem por um segmento de reta. Isto é importante, pois o cosseno do ângulo entre cada par de retas representa a correlação entre as respectivas variáveis. Quando o ângulo for reto, a correlação será nula, quando for zero, a correlação será 1, e quando for 180 graus, a correlação será -1.

Coeficientes de correlação entre as variáveis.

```
# Usando o pacote xtable para construir a tabela de correlação
print(xtable(cor(eur[, 3:5]), caption = "Matriz de correlação",
    label = "tab:biplot4"),
    caption.placement = "top", latex.environments = "flushleft",
        hline.after = c(-1,
        0))
```

Tabela 9.2: Matriz de correlação

	X1	X2	X3
X1	1.00	0.93	0.24
X2	0.93	1.00	0.21
X3	0.24	0.21	1.00

A matriz da tabela 9.2 mostra o que já vimos no gráfico. A correlação entre as variáveis X1 e X2 é igual a 0.93, uma correlação bem alta, enquanto a correlação entre X1 e X3 vale 0.24, e entre X2 e X3 vale 0.21. No gráfico 9.1, podemos ver perfeitamente, que o ângulo entre X1 e X2 é quase zero, enquanto os ângulos entre X1 e X3 e entre X2 e X3 são quase retos, ou seja, o poder de compra *per capita* em euros e o produto interno bruto(PIB) *per capita* são altamente relacionados, enquanto a taxa de inflação não tem muita relação com os dois primeiros indicadores.

As posições dos vários países na figura 9.1 mostram, pelos agrupamentos, que eles têm medidas parecidas nas variáveis. Luxemburgo está bem separado dos outros países, à direita quase em cima do eixo horizontal, na direção das variáveis X1 e X2. Isto mostra que Luxemburgo tem leituras altas nestas duas variáveis, ou seja, este país tem alto poder de compra e alto pib *per capita*. Por outro lado, Portugal está localizado do lado oposto às 3 variáveis, ou seja, este país tem valores baixos nas 3 variáveis o que realmente acontece; na tabela 9.1, vemos que Portugal tem, em relação aos outros países, um baixo poder de compra *per capita* em euros, um pequeno produto interno bruto (PIB) *per capita*, uma baixa taxa de inflação.

Os valores médios de X1, X2 e X3 são respectivamente, 20075,0; 124,6 e 3,47. A Dinamarca (DE) tem seus valores nas três variáveis bem perto destas médias e, à medida que os países vão se distanciando da origem, seus valores, ou pelo menos um valor, se afasta muito da média. A Bélgica (BE) tem uma alta taxa de inflação, porque se aproxima de X3, enquanto a Holanda (NE) tem uma baixa taxa de inflação, já que está oposta a X3.

9.2 Fibrose cística

A tabela 9.3 contém os dados sobre a função pulmonar em pacientes com fibrose cística, uma doença hereditária que afeta a capacidade pulmonar do doente.

```
# Lendo os dados de fibrose cística para um data.frame
fc ← read.table("C:/LivroExploratoria/fibcis.txt", header = T)
# Usando o pacote xtable para fazer a tabela de fibrose cística
print(xtable(fc, caption = "Fibrose cística", label = "tab:
    biplot5"), caption.placement = "top",
    latex.environments = "flushleft", hline.after = c(-1, 0))
```

As variáveis são as seguintes.

- id, idade.

Tabela 9.3: Fibrose cística

	id	sex	alt	pe	mc	vrf	vr	crf	cpt	pemax
1	7	0	109	13.10	68	32	258	183	137	95
2	7	1	112	12.90	65	19	449	245	134	85
3	8	0	124	14.10	64	22	441	268	147	100
4	8	1	125	16.20	67	41	234	146	124	85
5	8	0	127	21.50	93	52	202	131	104	95
6	9	0	130	17.50	68	44	308	155	118	80
7	11	1	139	30.70	89	28	305	179	119	65
8	12	1	150	28.40	69	18	369	198	103	110
9	12	0	146	25.10	67	24	312	194	128	70
10	13	1	155	31.50	68	23	413	225	136	95
11	13	0	156	39.90	89	39	206	142	95	110
12	14	1	153	42.10	90	26	253	191	121	90
13	14	0	160	45.60	93	45	174	139	108	100
14	15	1	158	51.20	93	45	158	124	90	80
15	16	1	160	35.90	66	31	302	133	101	134
16	17	1	153	34.80	70	29	204	118	120	134
17	17	0	174	44.70	70	49	187	104	103	165
18	17	1	176	60.10	92	29	188	129	130	120
19	17	0	171	42.60	69	38	172	130	103	130
20	19	1	156	37.20	72	21	216	119	81	85
21	19	0	174	54.60	86	37	184	118	101	85
22	20	0	178	64.00	86	34	225	148	135	160
23	23	0	180	73.80	97	57	171	108	98	165
24	23	0	175	51.10	71	33	224	131	113	95
25	23	0	179	71.50	95	52	225	127	101	195

- sex, sexo; 0-masculino, 1-feminino
- alt, altura em cm.
- pe, peso em kg.

- mc, massa corporal em percentual do normal.
- vrf, volume expiratório forçado.
- vr, volume residual.
- crf, capacidade residual funcional.
- cpt, capacidade pulmonar total.
- pemax, máxima pressão expiratória.

```
# Usando o pacote bpca para fazer o gráfico de biplot
par(mar = c(4, 4, 2, 1) + 0.1)
plot(bpca(fc), obj.cex = 0.5, var.cex = 0.5)
```

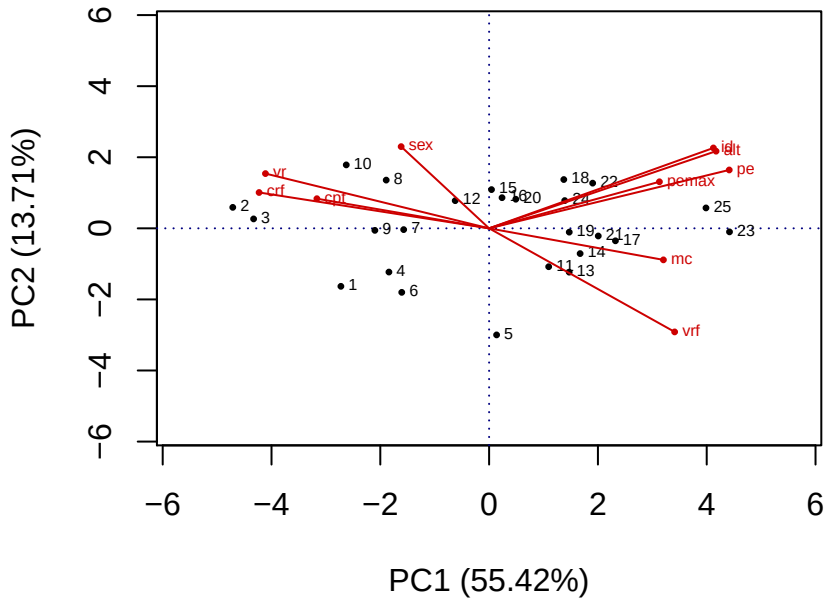


Figura 9.2: Biplot, resultado gráfico

Podemos ver, primeiramente, que as variáveis idade, altura, peso e máxima pressão expiratória são muito correlacionadas positivamente, isto porque a idade varia somente de 7 a 23 anos e, nestas idades, as 3 primeiras variáveis tem uma alta correlação positiva e a variável pemax, aumenta com a idade, mesmo para quem sofre da doença citada. Massa corporal e volume expiratório forçado também se correlacionam positivamente entre si e com as anteriores, porém a força da correlação é menor. O volume expiratório forçado tem uma correlação positiva e pequena com a variável pemax, o que talvez já seja um efeito da doença. A variável sexo tem uma correlação levemente negativa com as 3 primeiras variáveis e, talvez, na verdade, até seja não correlacionada com elas. Já com as variáveis mc e vrf tem uma correlação negativa e alta, mostrando que, quando mc e vrf crescem, o fator sexo decresce, ou seja, passa de 1 para 0, o que quer dizer que os homens têm valores maiores do que as mulheres naquelas 2 variáveis. As variáveis volume residual, capacidade residual funcional e capacidade pulmonar total são muito correlacionadas positivamente entre si, pois todas medem capacidade pulmonar, e estão correlacionadas positivamente com sexo, ou seja, os homens estão relacionados a altos valores destas medidas de capacidade pulmonar. Elas estão, ao mesmo tempo, muito correlacionadas negativamente com as outras variáveis situadas do lado esquerdo do gráfico. Isto mostra, principalmente com relação à idade, o avanço da doença afetando a capacidade pulmonar das pessoas afetadas.

Referências Bibliográficas

DAHL, D. B. *xtable: Export tables to LaTeX or HTML*. 2013. R package version 1.7-1. Disponível em: <http://CRAN.R-project.org/package=xtable>.

FARIA, J. C.; DEMETRIO, C. G. B. *bpca: Biplot of Multivariate Data Based on Principal Components Analysis.* Ilheus, Bahia, Brasil and Piracicaba, Sao Paulo, Brasil, 2013. Disponível em: <http://CRAN.R-project.org/package=bpca>.

FARIA, J. C.; GROSJEAN, P.; JELIHOVSCHI, E. *Tinn-R - GUI/Editor for R language and environment statistical computing.* 2013. Disponível em: <http://sourceforge.net/projects/tinn-r>.

FARIA, J. C.; JELIHOVSCHI, E. *fdth: Frequency Distribution Tables, Histograms and Poligons*. 2012. R package version 1.1-7. Disponível em: <http://CRAN.R-project.org/package=fdth>.

GENZ, A. et al. *mvtnorm: Multivariate Normal and t Distributions.* 2013. Disponível em: <http://CRAN.R-project.org/package=mvtnorm>.

GREENACRE, M. *Correspondence Analysis in Practice*. second. [S.l.]: chapman & Hall/CRC, 2007.

GREENACRE, M. *Biplots in Practice*. first. Bilbao, Spain: Fundacion BBVA, 2010.

NENADIC, O.; GREENACRE, M. Correspondence analysis in r, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, v. 20, n. 3, p. 1 – 13, 2007. Disponível em: <http://www.jstatsoft.org>.

RCORETEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>.

XIE, Y. *knitr: A general-purpose package for dynamic report generation in R*. 2013. Disponível em: <http://CRAN.R-project.org/package=knitr>.