

# Inferência Estatística com R

P. Maçaira, L. Bastos, S. Aguilar & I. Peres



# Contents

<b>Prefácio</b>	<b>5</b>
<b>1 Análise Descritiva</b>	<b>7</b>
1.1 Introdução . . . . .	7
1.2 Coleta e Armazenamento de Dados . . . . .	9
1.3 Tipos de Variáveis . . . . .	10
1.4 Variáveis Quantitativas . . . . .	11
1.5 Variáveis Qualitativas (ou categóricas) . . . . .	11
1.6 Medidas de Tendência Central . . . . .	12
1.7 Medidas de Variabilidade . . . . .	14
1.8 Medidas de Posição . . . . .	20
1.9 Referências . . . . .	23
<b>2 Análise Gráfica</b>	<b>25</b>
2.1 Introdução . . . . .	25
2.2 Variáveis Qualitativas - Nominiais e Ordinais . . . . .	25
2.3 Variáveis Quantitativas Discretas . . . . .	30
2.4 Variáveis Quantitativas Contínuas . . . . .	31
2.5 Histograma . . . . .	31
2.6 Boxplot . . . . .	35
2.7 Diagrama de Dispersão . . . . .	37
2.8 Séries Temporais . . . . .	41
2.9 Gráfico de Radar . . . . .	42



# Prefácio

Adicionar um vídeo nosso nos apresentando e o objetivo do livro.(pensar numa capa para o livro)

Sejam muito bem-vindos ao nosso livro on-line sobre Inferência Estatística.

Este livro destina-se a fornecer uma introdução abrangente aos métodos de inferência estatística e apresentar informações suficientes sobre cada método para que os leitores possam usá-los com sensatez. Desejamos que o leitor adquira o raciocínio necessário para, a partir dos dados, obter conclusões gerais acerca de uma população com base numa amostra.

O livro foi escrito para três públicos: (1) pessoas que se encontram utilizando estatística inferencial nos negócios quando podem não ter nenhum treinamento formal na área; (2) estudantes de graduação em Engenharia; (3) alunos de pós-graduação fazendo disciplina de estatística. Nós mesmos o usamos para alunos de pós-graduação e graduação da Pontifícia Universidade Católica do Rio de Janeiro, Brasil.

Para a maioria das seções, assumimos apenas que os leitores estão familiarizados com probabilidade estatística introdutória e com álgebra do ensino médio. Existem algumas seções que também exigem conhecimento de matrizes.

Usaremos o software R em todo o livro, um software gratuito e disponível em quase todos os sistemas operacionais. O R é uma ferramenta maravilhosa para todas as análises estatísticas e muito mais. Ao longo de todo o livro você se familiarizará com a linguagem e aprenderá a fazer inferência com o R.

A saída abaixo mostra as versões dos pacotes que usamos na compilação desta edição do livro. Alguns exemplos no livro não funcionarão com versões anteriores dos pacotes.

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --  
  
## v ggplot2 3.3.5      v purrr   0.3.4  
## v tibble  3.1.6      v dplyr  1.0.7  
## v tidyr   1.1.4      v stringr 1.4.0
```

```
## v readr    2.1.2      v forcats 0.5.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Finalmente, a saída lista alguns conflitos mostrando qual função será preferida quando uma função de mesmo nome estiver em vários pacotes.

O livro é diferente de outros livros de inferência estatística de várias maneiras.

1. É gratuito e online, tornando-o acessível a um grande público.
2. Ele usa R, que é um software gratuito, de código aberto e extremamente poderoso.
3. A versão online é continuamente atualizada. Você não precisa esperar até a próxima edição para que os erros sejam removidos ou novos métodos sejam discutidos. Atualizaremos o livro com frequência.
4. Existem dezenas de exemplos de dados reais retirados de nossa própria prática de consultoria.
5. Enfatizamos os métodos gráficos mais do que a maioria dos analistas. Usamos gráficos para explorar os dados, analisar a validade dos modelos ajustados e apresentar os resultados.

Boa leitura!

Paula Maçaira, Leonardo Bastos, Soraida Aguilar e Igor Peres.

Fevereiro 2022

Para citar a versão online deste livro, use o seguinte:

Maçaira, P.; Bastos, L.; Aguilar, S. & Peres, I. (2022) Inferência Estatística com R, 1ª edição, OTexts: Melbourne, Australia. OTexts.com/fpp3. Acessado em 2022-02-03.

# Chapter 1

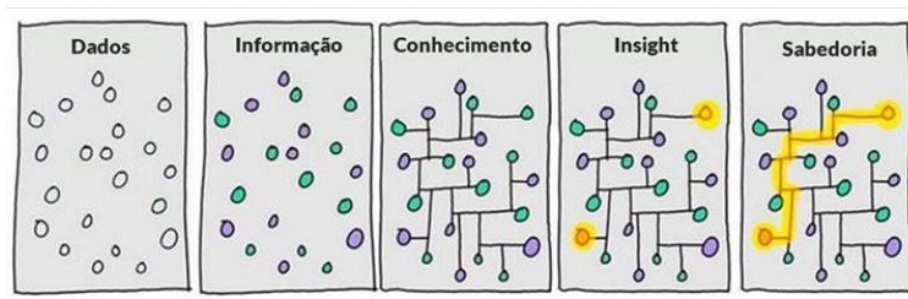
## Análise Descritiva

### 1.1 Introdução

A indústria 4.0, ou quarta revolução industrial, é a continuação do aperfeiçoamento das máquinas, um processo que começou com a primeira Revolução Industrial e nunca mais parou. Podemos dizer que a indústria 4.0 é a realidade na qual a tecnologia industrial está cada vez mais eficiente: mais inteligente, mais rápida e mais precisa. O termo é utilizado para caracterizar a utilização do que há de mais moderno para produzir bens de consumo: big data, internet das coisas, inteligência artificial e muito mais.

O big data é um conceito muito importante e que vem ganhando bastante notoriedade nos últimos anos. É o termo em Tecnologia da Informação (TI) que trata sobre grandes conjuntos de dados que precisam ser processados e armazenados. É um conceito-chave para a quarta revolução industrial, pois são esses dados que permitem às máquinas trabalharem com maior eficiência.

Dados sozinhos não podem dizer muita coisa. Por exemplo, eu posso te dizer que minha cor favorita é verde. Isso não te leva a tirar muitas conclusões sobre minha personalidade. Porém, se eu te der uma pequena tabela em excel com alguns dos meus livros favoritos, provavelmente você terá conclusões mais significativas sobre mim. O que aconteceu foi que você, munido da sua capacidade mental e analítica, aliada a uma certa quantidade de dados foi capaz de tirar alguns insights sobre um assunto. É exatamente aí que está a beleza e a grandiosidade da chamada “Era do Big Data”.



Big Data é basicamente análise de dados. De fato isso não é nenhuma novidade. Há muitos e muitos anos a humanidade coleta dados para serem analisados. A grande inovação está em aliar métodos antigos e limitados de análise de dados aos modernos recursos de hardware de alto processamento. Ou seja, agora é possível transitar todos esses cálculos e análises por meio de softwares desenvolvidos especificamente para trabalharem com enormes quantidades de dados. Uma solução de Big Data funciona com algoritmos complexos que trabalham a informação de modo a obter como saída os mais diversos tipos de insights.

A Análise Descritiva é a fase inicial deste processo de estudo dos dados coletados. Utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos. As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como porcentagens, índices e médias.

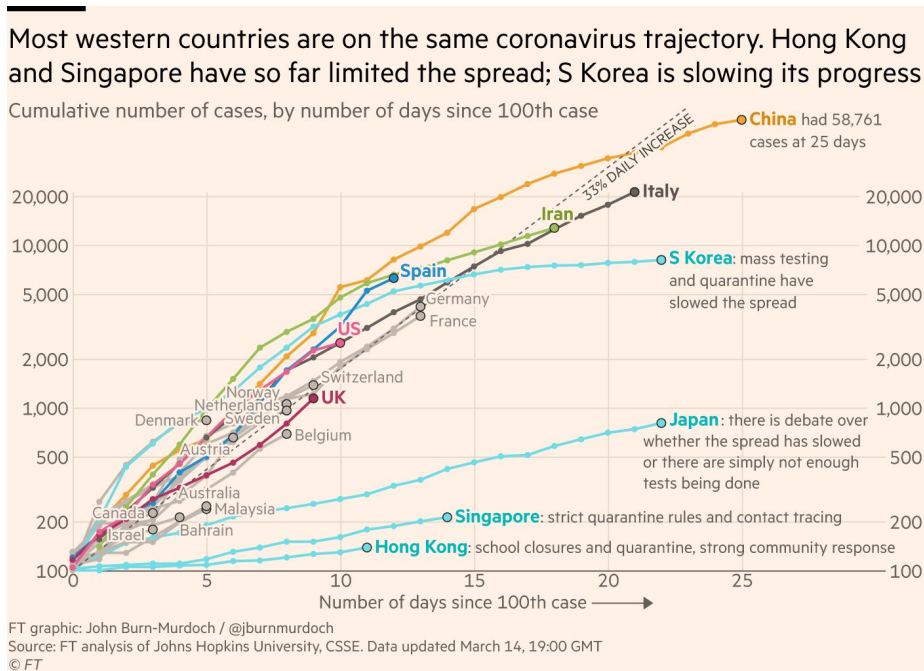
Ao se condensar os dados, perde-se informação, pois não se têm as observações originais. Entretanto, esta perda de informação é pequena se comparada ao ganho que se tem com a clareza da interpretação proporcionada.

A descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto.

Não só nos artigos técnicos direcionados para pesquisadores, mas também nos artigos de jornais e revistas escritos para o público leigo, é cada vez mais frequente a utilização destes recursos de descrição para complementar a apresentação de um fato, justificar ou referendar um argumento.

Durante a pandemia da COVID-19 foram publicadas ótimas visualizações de dados de casos e mortes, mas a mais conhecida delas é o gráfico de John Burn-Murdoch no Financial Times. Esta é uma ótima visualização e ajudou a apresentar gráficos em escala logarítmica para um público amplo.





Ao mesmo tempo em que o uso das ferramentas estatísticas vem crescendo, aumenta também o abuso de tais ferramentas. É muito comum vermos em jornais e revistas, até mesmo em periódicos científicos, gráficos - voluntariamente ou intencionalmente - enganosos e estatísticas obscuras para justificar argumentos polêmicos.

Neste capítulo iremos mergulhar no mundo da Análise de Descritiva de Dados, explorando estatísticas descritivas e análises gráficas com os pacotes mais atuais do software R [R Core Team, 2020].

## 1.2 Coleta e Armazenamento de Dados

### 1.2.1 Exemplo inicial

O primeiro caso da COVID-19 no Brasil foi confirmado em 25 de fevereiro de 2020, em São Paulo. Desde então o Governo Federal do Brasil reporta diariamente o número total de casos e mortes no país para cada Unidade da Federação, que corresponde aos 26 estados mais o Distrito Federal. No github do Wesley Cota [Cota, 2020] é possível extrair tais informações atualizadas para o dia anterior de forma sistematizada numa planilha de dados de maneira a realizar a entrada dos dados num programa de computador.

Para exemplificação, iremos extrair do github do Wesley Cota a planilha que possui informações agregadas para o Brasil e suas UFs. Este é o formato mais comum de uma base de dados, composta por linhas e colunas. Cada linha

contém os dados de uma Unidade da Federação (elemento) e as informações (variáveis) estão dispostas nas colunas. Dessa forma, a base de dados contém um número de linhas igual ao número de Unidades da Federação mais o total Brasil e um número de colunas igual ao número de variáveis sendo estudadas.

Estado	Casos	Mortes
TOTAL	25820745	629301
AC	102866	1877
AL	262163	6451
AM	537441	13981
AP	154329	2054
BA	1374028	28026
CE	1131297	25332
DF	617945	11197
ES	881190	13585
GO	1050619	25101
MA	389276	10501
MG	2777186	57440
MS	431152	9912
MT	639332	14342
PA	656272	17382
PB	504194	9738
PE	712083	20682
PI	347924	7405
PR	2015500	41320
RJ	1790994	69970
RN	431134	7758
RO	323699	6844
RR	143511	2099
RS	1850106	37000
SC	1458544	20664
SE	297332	6109
SP	4666941	158523
TO	273687	4008

Note que se usássemos uma base de dados que contém a evolução da doença por UF então teríamos uma combinação que seria uma linha para cada dia e cada estado, aumentando consideravelmente a dimensão da base.

### 1.3 Tipos de Variáveis

**Variável** é a característica de interesse que é medida em cada indivíduo da amostra ou população. Como o nome diz, seus valores variam de indivíduo para indivíduo. As variáveis podem ter valores numéricos ou não numéricos.

## 1.4 Variáveis Quantitativas

São as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser contínuas ou discretas.

- **Variáveis contínuas:** características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores não-inteiros (com casas decimais) fazem sentido. Usualmente devem ser medidas através de algum instrumento.
  - Exemplos: peso (balança), altura (régua), tempo (relógio), pressão arterial, idade.
- **Variáveis discretas:** características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente, são o resultado de contagens.
  - Exemplos: número de filhos, número de bactérias por litro de leite, número de casos de uma doença.

## 1.5 Variáveis Qualitativas (ou categóricas)

São as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.

- **Variáveis nominais:** não existe ordenação entre as categorias. Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.
- **Variáveis ordinais:** existe uma ordenação entre as categorias. Exemplos: escolaridade (1o, 2o, 3o graus), estágio da doença (inicial, intermediário, terminal), mês de observação (janeiro, fevereiro,..., dezembro).

Uma variável originalmente quantitativa pode ser coletada de forma qualitativa. Por exemplo, a variável idade, medida em anos completos, é quantitativa (contínua); mas, se for informada apenas a faixa etária (0 a 5 anos, 6 a 10 anos etc.), é qualitativa (ordinal). Outro exemplo é o peso dos lutadores de boxe, uma variável quantitativa (contínua) se trabalhamos com o valor obtido na balança, mas qualitativa (ordinal) se o classificarmos nas categorias do boxe (peso-pena, peso-leve, peso-pesado etc.).

Outro ponto importante é que nem sempre uma variável representada por números é quantitativa. O número do telefone de uma pessoa, o número da casa, o número de sua identidade. Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável sexo passou a ser quantitativa.

## 1.6 Medidas de Tendência Central

A tendência central de uma variável em um conjunto de dados é caracterizada pelo valor típico dessa variável. Essa é uma maneira de resumir a informação contida nos dados, pois escolheremos um valor para representar todos os outros. Assim, poderíamos perguntar, por exemplo, qual é a altura típica dos brasileiros adultos no final da década de 90 e compará-la com o valor típico da altura dos brasileiros no final da década de 80, a fim de verificar se os brasileiros estão se tornando, em geral, mais altos, mais baixos ou não sofreram nenhuma alteração em sua altura típica. Fazer essa comparação utilizando medidas-resumo (as alturas típicas em cada período) é bem mais sensato do que comparar os dois conjuntos de dados valor a valor, o que seria inviável. Mas, como identificar o valor típico de um conjunto de dados?

Existem três medidas que podem ser utilizadas para descrever a tendência central de um conjunto de dados: a média, a mediana e a moda. Apresentaremos essas três medidas e discutiremos suas vantagens e desvantagens.

### 1.6.1 Média Aritmética Simples

A média aritmética simples (que chamaremos apenas de média) é a medida de tendência central mais conhecida e usada para o resumo de dados. Essa popularidade pode ser devida à facilidade de cálculo e à idéia simples que ela nos sugere. De fato, se queremos um valor que represente a altura dos brasileiros adultos, por que não medir as alturas de uma amostra de brasileiros adultos, somar os valores e dividir esse “bolo” igualmente entre os participantes? Essa é a idéia da média aritmética.

Para apresentar a média, primeiramente vamos definir alguma notação. A princípio, essa notação pode parecer desnecessária, mas facilitará bastante nosso trabalho futuro.

$$\text{Notação} \left\{ \begin{array}{ll} n & \text{tamanho da amostra} \\ x_i & \text{valor da } i\text{-ésima observação} \\ \sum_{i=1}^n x & \text{soma de todas as observações} \\ \bar{x} & \text{símbolo que representa a média aritmética simples} \end{array} \right.$$

Assim,

$$\bar{x} = \frac{\text{soma de todas as observações}}{n} = \frac{\sum_{i=1}^n x}{n}$$

Exemplo: No conjunto de dados (1.3, 0.7, 5.8, 2.4, 1.2), temos  $n = 5$ ,  $x_1 = 1.3$ ,  $x_2 = 0.7$ ,  $x_3 = 5.8$ ,  $x_4 = 2.4$  e  $x_5 = 1.2$ , portanto  $\sum_{i=1}^5 x_i = 1.3 + 0.7 + 5.8 + 2.4 + 1.2 = 11.4$  e assim  $\bar{x} = \frac{11.4}{5} = 2.28$ .

Se esses seis valores representassem, por exemplo, as quantidades de peixe pescado (em toneladas) durante cinco dias da semana, a quantidade típica pescada por dia, naquela semana, seria 2,28 toneladas. Como estamos representando o valor típico pela média aritmética, podemos falar em quantidade média diária naquela semana.

Fazendo no R:

```
mean(c(1.3, 0.7, 5.8, 2.4, 1.2))
```

```
## [1] 2.28
```

### 1.6.2 Mediana

A mediana de um conjunto de dados é definida como sendo o “valor do meio” desse conjunto de dados, dispostos em ordem crescente, deixando metade dos valores acima dela e metade dos valores abaixo dela.

Como calcular a mediana? Basta seguir sua definição. Vejamos:

- **$n$  é ímpar:** Existe apenas um “valor do meio”, que é a mediana.
  - Seja o conjunto de dados (1.3, 0.7, 5.8, 2.4, 1.2).
  - Ordenando os valores (0.7, 1.2, 1.3, 2.4, 5.8).
  - O valor do meio é o 1.3.
  - A mediana é o valor 1.3.
- **$n$  é par:** Existem dois “valores do meio”. A mediana é a média aritmética simples deles.
  - Seja o conjunto de dados (1.3, 0.7, 5.8, 2.4, 1.2, 2.1).
  - Ordenando os valores (0.7, 1.2, 1.3, 2.1, 2.4, 5.8).
  - Os valores do meio são 1.3 e 2.1.
  - A mediana é  $(1.3+2.1)/2=1.7$ .

Fazendo no R:

```
median(c(1.3, 0.7, 5.8, 2.4, 1.2))
```

```
## [1] 1.3
```

```
median(c(1.3, 0.7, 5.8, 2.4, 1.2, 2.1))
```

```
## [1] 1.7
```

Como medida de tendência central, a mediana é até mais intuitiva do que a média, pois representa, de fato, o centro (meio) do conjunto de valores ordenados. Assim como a média, o valor da mediana não precisa coincidir com algum dos valores do conjunto de dados. Em particular, quando os dados forem de natureza contínua, essa coincidência dificilmente ocorrerá.

### 1.6.3 Moda

Uma maneira alternativa de representar o que é “típico” é através do valor mais frequente da variável, chamado de moda.

Como calcular a moda? Basta verificar o valor que “aparece” mais vezes. Vejamos:

- No conjunto de dados (1, 2, 3, 3, 4, 5, 5, 5, 5, 5), há apenas uma moda, o valor 5, portanto o conjunto de dados é **unimodal**.
- No conjunto de dados (1, 2, 2, 2, 2, 3, 4, 5, 6, 6, 6, 6, 7, 9), existem duas modas, os valores 2 e 6, portanto o conjunto de dados é **bimodal**.
- Nem sempre a moda existe ou faz sentido, no conjunto de dados (1, 2, 3, 4, 5, 6, 7, 8, 9), não existe um valor mais frequente que os demais, portanto o conjunto de dados é **amodal**.

Para usar a função que calcula a moda (`Mode`) no R temos que instalar e carregar o pacote `pracma`:

```
library(pracma)
Mode(c(1,2,3,3,4,5,5,5,5,5))

## [1] 5
Mode(c(1,2,2,2,2,3,4,5,6,6,6,6,7,9)) # escolherá o menor valor, caso haja empate

## [1] 2
Mode(c(1,2,3,4,5,6,7,8,9))

## [1] 1
```

A moda é também a única das medidas de tendência central que faz sentido no caso de variáveis qualitativas. Assim, a categoria dessas variáveis que aparecer com maior frequência é chamada de categoria modal.

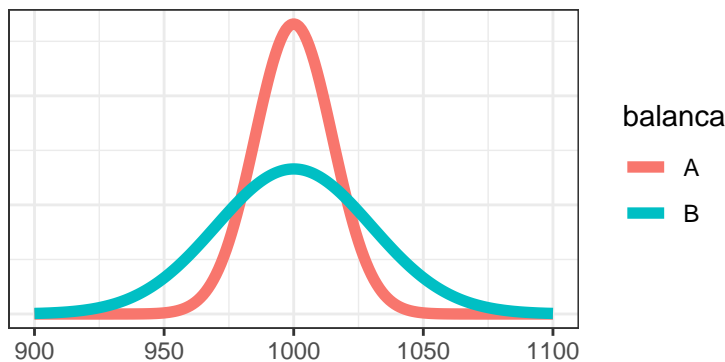
## 1.7 Medidas de Variabilidade

As medidas de tendência central (média, mediana, moda) conseguem resumir em um único número, o valor que é “típico” no conjunto de dados. Mas, será que, somente com essas medidas, conseguimos descrever adequadamente o que ocorre em um conjunto de dados?

Vejamos um exemplo: quando pesamos algo em uma balança, esperamos que ela nos dê o verdadeiro peso daquilo que estamos pesando. No entanto, se fizermos várias medições do peso de um mesmo objeto em uma mesma balança, teremos diferentes valores para o peso deste objeto. Ou seja, existe variabilidade nas medições de peso fornecidas pela balança. Neste caso, quanto menor a variabilidade desses valores, mais precisa é a balança (considerando que a média das medidas de peso coincida como seu valor real). Observe na Figura abaixo,

onde estão representadas as distribuições das medições do peso de uma esfera de 1000g, feitas por duas balanças (A e B). As duas balanças registram o mesmo peso médio de 1000g (média dos pesos de todas as medições feitas). Isto é, as duas balanças tipicamente acertam o verdadeiro peso da esfera. Porém, pela Figura, podemos notar que

- as medições da balança A *variam pouco* em torno de 1000g: oscilam bastante entre cerca de 950g e 1050g (uma “imprecisão” de 50g)
- as medições da balança B *variam muito* em torno de 1000g: oscilam basicamente entre 900g e 1100g (uma “imprecisão” de 100g)



Dois conjuntos de dados podem ter a mesma medida de centro (valor típico), porém uma dispersão diferente em torno desse valor. Desse modo, além de uma medida que nos diga qual é o valor “típico” do conjunto de dados, precisamos de uma medida do grau de dispersão (variabilidade) dos dados em torno do valor típico.

O objetivo das medidas de variabilidade é quantificar esse grau de dispersão. Nesta seção, apresentaremos três dessas medidas (amplitude total, desvio-padrão e coeficiente de variação), discutindo suas vantagens e desvantagens. Em discussões posteriores, apresentaremos medidas de variabilidade alternativas.

### 1.7.1 Amplitude Total

A medida de variabilidade mais simples é a chamada amplitude total (AT), que é a diferença entre o valor máximo e o valor mínimo de um conjunto de dados.

$$AT = \text{Máximo} - \text{Mínimo}$$

Exemplo: Medição do peso de uma esfera de 1000g em duas balanças (A e B).

Estatísticas	Balança A	Balança B
Mínimo	945g	895g

Estatísticas	Balança A	Balança B
Máximo	1040g	1095g
AT	1040-945=95g	1095-895=200g

A variabilidade das medições de peso da balança B é maior que a variabilidade das medições de peso da balança A (apesar do valor médio ser igual).

Embora seja uma medida simples de variabilidade, a amplitude total é um tanto grosseira, pois depende somente de dois valores do conjunto de dados (máximo e mínimo), não captando o que ocorre com os outros valores.

Fazendo no R:

```
1040-945
```

```
## [1] 95
```

```
1095-895
```

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

### 1.7.2 Desvio Padrão

Uma boa medida de dispersão deve considerar todos os valores do conjunto de dados e resumir o grau de dispersão desses valores em torno do valor típico.

Considerando a média como a medida de tendência central, podemos pensar em medir a dispersão (desvio) de cada valor do conjunto de dados em relação à ela. A medida mais simples de desvio entre duas quantidades é a diferença entre elas. Assim, para cada valor  $x_i$ , teremos o seu desvio em relação à  $\bar{x}$  representado por  $(x_i - \bar{x})$ .

Exemplo: No conjunto de dados 1, 1, 2, 3, 4, 4, 5, 6, 7, 7, relativo ao número de filhos de 10 mulheres, temos  $\bar{x} = 4$  filhos. Na tabela abaixo, a coluna 1 mostra esses 10 valores e a coluna 2 mostra o desvio de cada um deles até a média.

Coluna 1 (Xi)	Coluna 2 (Xi-Media)	Coluna 3 (Xi-Media) <sup>2</sup>
1	-3	9
1	-3	9
2	-2	4
3	-1	1
4	0	0
4	0	0
5	1	1
6	2	4
7	3	9
7	3	9
Soma	0	46



### A ideia do desvio padrão

Como temos um desvio para cada elemento, poderíamos pensar em resumi-los em um desvio típico, a exemplo do que fizemos com a média. Porém, quando somarmos esses desvios para o cálculo do desvio médio, a soma dará sempre zero, como pode ser visto na coluna 2 do exemplo anterior. Isto ocorre com qualquer conjunto de dados, pois os desvios negativos sempre compensaram os positivos.

No entanto, os sinais dos desvios não são importantes para nossa medida de dispersão, já que estamos interessados na quantidade de dispersão e não na direção dela. Portanto, eliminaremos os sinais elevando os desvios ao quadrado, como mostrado na coluna 3. A soma desses desvios ao quadrado pode ser, então, dividida entre os participantes do “bolo”. Na verdade, por razões absolutamente teóricas, dividiremos essa soma pelo total de participantes menos 1 ( $n - 1$ ). Assim, usando a notação definida anteriormente, teremos

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

Para os dados do exemplo, teremos  $46/(10-1) = 5,11$ . Esse valor pode ser visto como uma quase-média dos desvios ao quadrado e é chamado de **variância**.

A variância seria nossa medida de variabilidade se não fosse o fato de que ela está expressa em uma unidade diferente da unidade dos dados, pois, ao elevarmos os desvios ao quadrado, elevamos também as unidades de medida em que eles estão expressos. No caso dos dados do exemplo, medidos em número de filhos, a variância vale 5,11 “filhos ao quadrado”, algo que não faz nenhum sentido.

Para eliminar esse problema, extraímos a raiz quadrada da variância e, finalmente, temos a nossa medida de variabilidade, que chamaremos desvio-padrão (DP).

$$DP = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

O desvio-padrão, como o nome já diz, representa o desvio típico dos dados em relação à média, escolhida como medida de tendência central. No exemplo, temos que o desvio padrão vale 2,26. Isto significa que a distância típica (padrão) de cada mãe até o número médio de filhos (4 filhos) é de 2,26 filhos. Quanto maior o desvio-padrão, mais diferentes entre si serão as quantidades de filhos de cada mãe.

O desvio-padrão, em alguns livros chamado de  $s$ , é uma medida sempre positiva. Se observarmos a maneira como ele é calculado, veremos que não há como obter um valor negativo.

Fazendo no R:

```
round(var(c(1,1,2,3,4,4,5,6,7,7)),2) # função round() serve para arredondar o número
```

```
## [1] 5.11
```

```
round(sd(c(1,1,2,3,4,4,5,6,7,7)),2)
```

```
## [1] 2.26
```

Exemplo: Os agentes de fiscalização de certo município realizam, periodicamente, uma vistoria nos bares e restaurantes para apurar possíveis irregularidades na venda de seus produtos. A seguir, são apresentados dados de uma vistoria sobre os pesos (em gramas) de uma amostra de 10 bifes, constantes de um cardápio de um restaurante como “bife de 200 gramas”.

```
## [1] 170 175 180 185 190 195 200 200 200 205
```

Como podemos notar, nem todos os “bifes de 200 gramas” pesam realmente 200 gramas. Esta variação é natural e é devida ao processo de produção dos bifes. No entanto, esses bifes deveriam pesar cerca de 200 gramas e com pouca variação em torno desse valor. Com o auxílio do R, calcularemos a média e o desvio-padrão.

```
mean(c(170,175,180,185,190,195,200,200,200,205))
```

```
## [1] 190
```

```
sd(c(170,175,180,185,190,195,200,200,200,205))
```

```
## [1] 12.0185
```

Os bifes desse restaurante pesam, em média, 190 gramas, com um desvio-padrão de 12 gramas. Ou seja, os pesos dos “bifes de 200 gramas” variam tipicamente entre 178 e 202 gramas. Analisando esses valores, concluímos que esse restaurante pode estar lesando a maior parte de seus clientes.

Para casos como esse, os agentes fiscalizadores podem estabelecer parâmetros (valores) para saber até quanto a média pode se desviar do valor correto e o quanto de variação eles podem permitir numa amostra para concluir que o processo de produção de bifes não possui problemas. Por exemplo, a média da amostra não poderia ser inferior a 198 gramas, com um desvio-padrão que não seja superior a 5% dessa média.

Essas idéias são utilizadas no controle do processo de produção das indústrias, onde já se espera alguma variação entre as unidades produzidas. Porém, essa variação deve estar sob controle. Numa indústria farmacêutica, por exemplo, espera-se que os comprimidos de uma certa droga sejam produzidos com uma certa variação em sua composição (maior ou menor quantidade do princípio ativo), devido à própria maneira como os comprimidos são produzidos (máquinas, pessoas etc.). No entanto, esta variação deve ser pequena, para que não sejam produzidos comprimidos inócuos (com pouco do princípio ativo) ou

com extra-dosagem do princípio ativo, o que, em ambos os casos, pode causar sérias complicações à saúde do paciente.

O desvio-padrão nos permite distinguir numericamente conjuntos de dados de mesmo tamanho, mesma média, mas que são visivelmente diferentes. Usando o desvio-padrão, também conseguimos representar numericamente a variabilidade das medições das balanças A e B, que, apesar de possuírem a mesma média, possuem variabilidades bastante diferentes.

Quando os conjuntos de dados a serem comparados possuem médias diferentes, a comparação da variabilidade desses conjuntos deve levar em conta essa diferença. Por esta e outras razões, definiremos uma terceira medida de variabilidade, o **coeficiente de variação**.

### 1.7.3 Coeficiente de Variação

Ao analisarmos o grau de dispersão de um conjunto de dados, poderemos nos deparar com uma questão do tipo: um desvio-padrão de 10 unidades é pequeno ou grande?

Vejamos:

- Se estivermos trabalhando com um conjunto de dados cuja média é 10.000, um desvio típico de 10 unidades em torno dessa média significa pouca dispersão;
- Mas, se a média for igual a 100, um desvio típico de 10 unidades em torno dessa média significa muita dispersão.

Assim, antes de responder se um desvio-padrão de 10 unidades é grande ou pequeno, devemos avaliar sua magnitude em relação à média:

- No primeiro caso, o desvio-padrão corresponde a 0,1% da média
- No segundo caso, o desvio-padrão corresponde a 10% da média

À essa razão entre o desvio-padrão e a média damos o nome de **Coeficiente de Variação**:

$$CV = \frac{\text{Desvio Padrão}}{\text{Média}}$$

Quanto menor o Coeficiente de Variação de um conjunto de dados, menor é a sua variabilidade. O Coeficiente de Variação expressa o quanto da escala de medida, representada pela média, é ocupada pelo desvio-padrão.

O Coeficiente de Variação é uma medida adimensional, isto é, não depende da unidade de medida. Essa característica nos permite usá-lo para comparar a variabilidade de conjuntos de dados medidos em unidades diferentes, o que seria impossível usando o desvio-padrão.

Exemplo: Numa pesquisa na área de Saúde Ocupacional, deseja-se comparar a idade de motoristas e cobradores de ônibus da região metropolitana de Belo Horizonte. Algumas estatísticas descritivas são apresentadas na Tabela abaixo.

Grupo	n	Media	DP	CV
Motoristas	150	35.6	5.08	0.143
Cobradores	50	22.6	3.11	0.137

Os motoristas são, em média, 13 anos mais velhos do que os cobradores. Ao compararmos o grau de dispersão dos dois grupos usando o desvio-padrão, concluiríamos que os motoristas são menos homogêneos quanto à idade do que os cobradores. Ao fazermos isso, estamos esquecendo que, apesar de estarem em unidades iguais, as medidas de idade nos dois grupos variam em escalas diferentes. As idades dos motoristas variam em torno dos 35 anos e podem chegar até 18 anos (idade mínima para se conseguir a habilitação), numa amplitude de 17 unidades. Enquanto isso, as idades dos cobradores variam em torno de 22 anos e também só podem chegar até a 18 anos, uma amplitude de apenas 4 anos. Assim, os motoristas tem a possibilidade de ter um desvio-padrão maior do que o dos cobradores. Se levarmos em conta a escala de medida, usando o coeficiente de variação, veremos que os motoristas são somente um pouco mais heterogêneos (dispersos) quanto à idade do que os cobradores.

Fazendo no R:

```
media=c(35.6,22.6)
desvio=c(5.08,3.11)
round(desvio/media,3)
```

```
## [1] 0.143 0.138
```

## 1.8 Medidas de Posição

Quando falamos de posição ou colocação de um indivíduo em uma corrida ou em um teste como o Vestibular, frequentemente nos referimos ao seu posto, como 1º, 2º, 3º, 29º ou último lugar. Mas, para sabermos se uma dada colocação é ou não um bom resultado, precisamos informar quantos indivíduos participaram da corrida ou do Vestibular.

A medida de posição que veremos aqui, os percentis, solucionam este e outros problemas de posicionamento (ranking). A posição de um indivíduo no conjunto de dados é mostrada, pelo percentil, contando-se (em porcentagem) quantos indivíduos do conjunto têm valores menores que o deste indivíduo.

Como veremos, esta medida de posição pode ser usada para comparar a posição do indivíduo em diferentes conjuntos de dados, nos quais foram medidas as mesmas variáveis ou variáveis diferentes.

### 1.8.1 Percentis

Considere o trecho a seguir, sobre a posição do Brasil, entre os países do mundo, quanto à renda per capita:

*O Brasil obviamente não é país rico, mas também não está entre os mais pobres. ... Mais de três quartos da população mundial vivem em países de renda per capita menor*

Neste caso, a posição do Brasil é dada pela quantidade de países que têm renda per capita menores que o Brasil, a saber três quartos ou 75%. O mesmo tipo de raciocínio fazemos quando dizemos que certo aluno está entre os 5% melhores do colégio. Não precisamos nem saber quantos alunos tem o colégio ou em quantos países estão sendo consideradas as rendas. Aqui já houve uma padronização da posição usando-se a porcentagem de alunos ou países com desempenho ou renda abaixo do valor considerado. É este raciocínio que define os percentis.

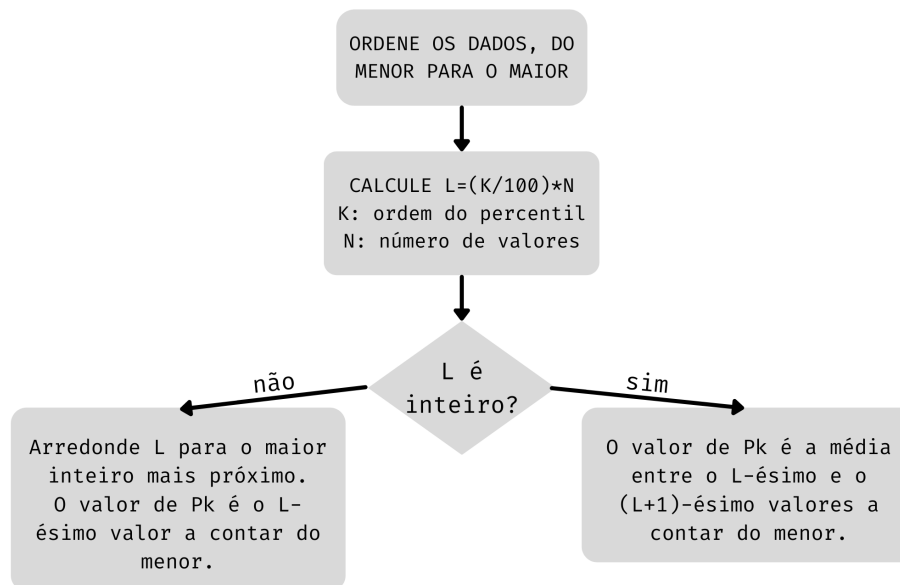
**O percentil de ordem  $K$  (onde  $k$  é qualquer valor entre 0 e 100), denotado por  $P_k$ , é o valor tal que  $K\%$  dos valores do conjunto de dados são menores ou iguais a ele.**

Assim, o percentil de ordem 10, o  $P_{10}$ , é o valor da variável tal que 10% dos valores são menores ou iguais a ele; o percentil de ordem 65 deixa 65% dos dados menores ou iguais a ele etc.

Os percentis de ordem 10, 20, 30, ..., 90 dividem o conjunto de dados em dez partes com mesmo número de observações e são chamados de **decis**.

Os percentis de ordem 25, 50 e 75 dividem o conjunto de dados em quatro partes com o mesmo número de observações. Assim, estes três percentis recebem o nome de quartis – **primeiro quartil (Q1), segundo quartil (Q2) e terceiro quartil (Q3)**, respectivamente. O segundo quartil é a já conhecida mediana.

Existem vários processos para calcular os percentis, usando interpolação. Vamos ficar com um método mais simples, mostrado na Figura a seguir. As diferenças serão muito pequenas e desaparecerão à medida que aumenta o número de dados.



Considere as notas finais dos 40 candidatos ao curso de Direito no Vestibular de certa faculdade, já colocadas em ordem crescente:

```
## [1] 40 41 42 42 44 47 48 48 49 49 51 52 53 58 59 62 63 64 65 66 67 68 69 70 75
## [26] 76 83 83 85 86 86 87 87 88 92 93 94 95 97 98
```

Vamos calcular alguns percentis:

- Percentil de ordem 10: 10% de 40 = 4. Então o  $P_{10}$  = média(4º e 5º valores) =  $(42 + 44)/2 = 43$ .
- Percentil de ordem 95: 95% de 40 = 38. Então o  $P_{95}$  = média(38º e 39º valores) =  $(95 + 97)/2 = 96$ .
- Primeiro Quartil: 25% de 40 = 10. Então o  $Q_1$  = média(10º e 11º valores) =  $(49 + 51)/2 = 50$ .
- Terceiro Quartil: 75% de 40 = 30. Então o  $Q_3$  = média(30º e 31º valores) =  $(86 + 86)/2 = 86$ .
- Mediana: 50% de 40 = 20. Então mediana = média(20º e 21º valores) =  $(66 + 67)/2 = 66,5$ .

Fazendo no R:

```
dados_ex7=c(40,41,42,42,44,47,48,48,49,
            49,51,52,53,58,59,62,63,64,
            65,66,67,68,69,70,75,76,83,
```

```
83,85,86,86,87,87,88,92,93,94,95,97,98)
quantile(dados,probs=c(0.1,0.95,0.25,0.75,0.5))
```

```
##      10%      95%      25%      75%      50%
## 174.50 202.75 181.25 200.00 192.50
```

Veja que os valores encontrados pelo R não coincidem com os calculados “na mão”, isto porque, o R utiliza um método de interpolação para calcular valores que não estão presentes na amostra.

## 1.9 Referências





## Chapter 2

# Análise Gráfica

### 2.1 Introdução

Já sabemos que as variáveis de um estudo dividem-se em quatro tipos:

- Qualitativas:
  - nominais
  - ordinais
- Quantitativas:
  - discretas
  - contínuas

Os dados gerados por esses tipos de variáveis são de naturezas diferentes e devem receber tratamentos diferentes. Portanto, vamos estudar as ferramentas mais adequadas para cada tipo de dados, separadamente.

### 2.2 Variáveis Qualitativas - Nominais e Ordinais

Na base de dados `mtcars`, uma das duas variáveis qualitativas presentes é a categoria da transmissão (automática ou manual). Para organizar os dados provenientes de uma variável qualitativa, é usual fazer uma tabela de frequências, como a tabela abaixo, onde estão apresentadas as frequências com que ocorre cada um dos tipos de transmissão no total dos 32 carros observados. Cada categoria da variável transmissão (automática, manual) é representada numa linha da tabela. Há uma coluna com as contagens de carros em cada categoria (frequência absoluta) e outra com os percentuais que essas contagens representam no total de carros (frequência relativa). Esse tipo de tabela representa a distribuição de frequências dos carros segundo a variável transmissão.

Como a variável transmissão é qualitativa nominal, ou seja, não há uma ordem natural em suas categorias, a ordem das linhas da tabela pode ser qualquer uma.

É comum a disposição das linhas pela ordem decrescente das frequências das classes.

Transmissão	Frequência Absoluta	Frequência Relativa (%)
Automático	19	59.38
Manual	13	40.62
Total	32	100.00

Quando a variável tabelada for do tipo qualitativa ordinal, as linhas da tabela de frequências devem ser dispostas na ordem existente para as categorias. A tabela abaixo mostra a distribuição de frequências das coletas segundo o mês de observação na base de dados *airquality*, que é uma variável qualitativa ordinal. Nesse caso, podemos acrescentar mais duas colunas com as frequências acumuladas (absoluta e relativa), que mostram, para cada mês, a frequência das coletas até aquele mês. Por exemplo, até o mês de julho, foram coletadas 92 amostras, o que representa 60,13% do total de amostras coletadas.

Transmissão	Frequência Absoluta	Frequência Relativa (%)	Frequência Absoluta Acumulada	Frequência Relativa Acumulada (%)
5	31	20.26	31	20.26
6	30	19.61	61	39.87
7	31	20.26	92	60.13
8	31	20.26	123	80.39
9	30	19.61	153	100
Total	153	100.00	-	-

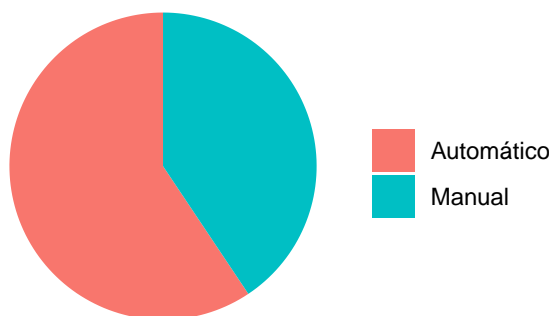
Note que as frequências acumuladas não fazem sentido em distribuição de frequências de variáveis para as quais não existe uma ordem natural nas categorias, as qualitativas nominais.

A visualização da distribuição de frequências de uma variável fica mais fácil se fizermos um gráfico a partir da tabela de frequências. Existem vários tipos de gráficos, dependendo do tipo de variável a ser representada. Para as variáveis do tipo qualitativas, abordaremos dois tipos de gráficos: os de setores e os de barras.

Os gráficos de setores, mais conhecidos como gráficos de pizza ou torta, são construídos dividindo-se um círculo (pizza) em setores (fatias), um para cada categoria, que serão proporcionais à frequência daquela categoria.

A Figura a seguir mostra um gráfico de setores para a variável transmissão, construído a partir da primeira tabela de frequências. Através desse gráfico, fica mais fácil perceber que os carros automáticos são a grande maioria dos carros. Como esse gráfico contém todas as informações da tabela, pode substituí-la com a vantagem de tornar análise dessa variável mais agradável.

```
library(ggplot2)
df = data.frame(table(mtcars$am))
levels(df$Var1) = c("Automático","Manual")
ggplot(df, aes(x="", y=Freq, fill=Var1)) +
  geom_bar(stat="identity") +
  coord_polar("y", start=0) +
  theme_void() +
  theme(legend.title = element_blank())
```

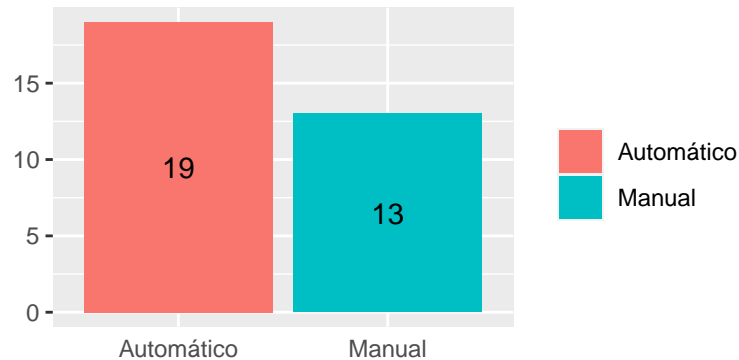


Quando houver mais de duas categorias de uma variável nominal, a disposição no gráfico de setores deve ser pela ordem decrescente das frequências, no sentido horário. A categoria “outros”, quando existir, deve ser sempre a última, mesmo não seja a de menor frequência.

As vantagens da representação gráfica das distribuições de frequências ficam ainda mais evidentes quando há a necessidade de comparar vários grupos com relação à variáveis que possuem muitas categorias, como veremos mais adiante.

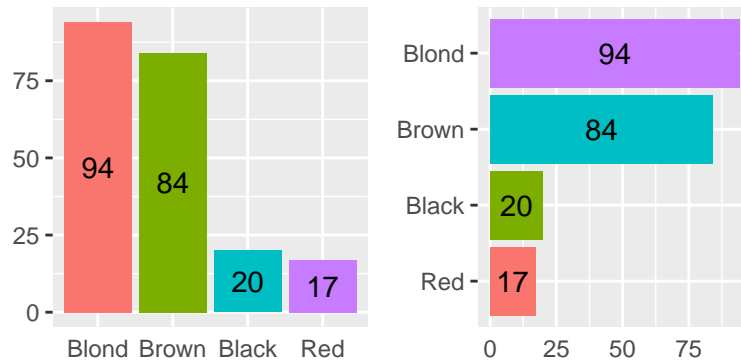
Uma alternativa ao gráfico de setores é o gráfico de barras (colunas) como o da figura a seguir. Ao invés de dividirmos um círculo, dividimos uma barra. Note que, em ambos os gráficos, as frequências relativas das categorias devem somar 100%. Aliás, esse é a idéia dos gráficos: mostrar como se dá a divisão (distribuição) do total de elementos (100%) em partes (fatias).

```
df = data.frame(table(mtcars$am))
levels(df$Var1) = c("Automático","Manual")
ggplot(df, aes(x=Var1, y=Freq, fill=Var1)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = Freq), position=position_stack(vjust = 0.5)) +
  theme(legend.title = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank())
```



Uma situação diferente ocorre quando desejamos comparar a distribuição de frequências de uma mesma variável em vários grupos, como por exemplo, a frequência de estudantes com olhos azuis entre todas as cores de cabelo na base de dados `HairEyeColor`. Se quisermos usar o gráfico de setores para fazer essa comparação, devemos fazer quatro gráficos, um para cada cor de cabelo, com duas fatias cada um (olhos azuis e olhos não azuis). Uma alternativa é a construção de um gráfico de colunas (barras) como os gráficos das figuras a seguir, onde há uma barra para cada cor de cabelo representando a frequência de estudantes com olhos azuis e aquela cor de cabelo. Além de economizar espaço na apresentação, permite que as comparações sejam feitas de maneira mais rápida.

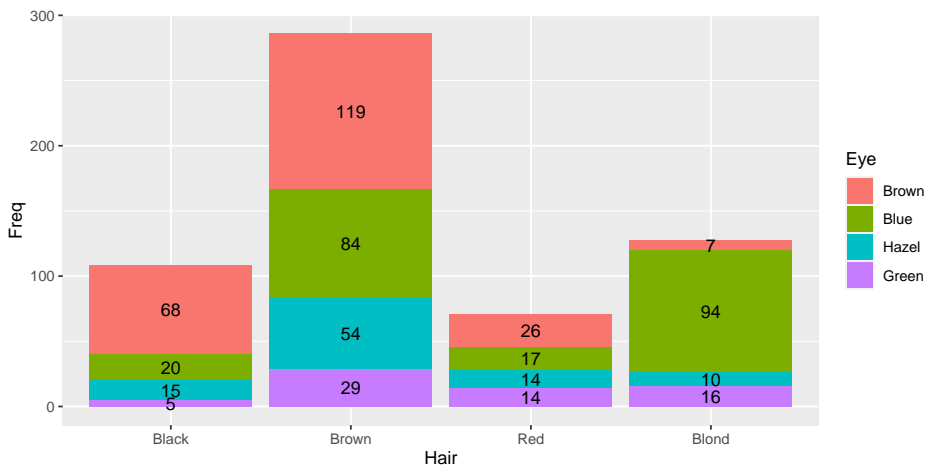
```
df = data.frame(HairEyeColor[,1] + HairEyeColor[,2])
df = df %>% filter(Eye == "Blue")
p1 = df %>% mutate(name = fct_reorder(Hair, desc(Freq))) %>%
  ggplot(aes(x=name, y=Freq, fill=name)) +
  geom_bar(position="dodge", stat="identity") +
  geom_text(aes(label = Freq), position=position_stack(vjust = 0.5)) +
  theme(legend.position = "none",
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank())
p2 = df %>% mutate(name = fct_reorder(Hair, Freq)) %>%
  ggplot(aes(x=name, y=Freq, fill=name)) +
  geom_bar(position="dodge", stat="identity") +
  geom_text(aes(label = Freq), position=position_stack(vjust = 0.5)) +
  coord_flip() +
  theme(legend.position = "none",
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank())
gridExtra::grid.arrange(p1,p2,ncol=2)
```



A ordem dos grupos pode ser qualquer, ou aquela mais adequada para a presente análise. Frequentemente, encontramos as barras em ordem decrescente, já antecipando nossa intuição de ordenar os grupos de acordo com sua frequência para facilitar as comparações. Caso a variável fosse do tipo ordinal, a ordem das barras seria a ordem natural das categorias, como na tabela de frequências.

A figura abaixo mostra um gráfico de barras que pode ser usado da comparação da distribuição de frequências de uma mesma variável em vários grupos. É também uma alternativa ao uso de vários gráficos de setores, sendo, na verdade, a junção de dois gráficos com os das figuras acima num só gráfico. Porém, esse tipo de gráfico só deve ser usado quando não houver muitos grupos a serem comparados e a variável em estudo não tiver muitas categorias.

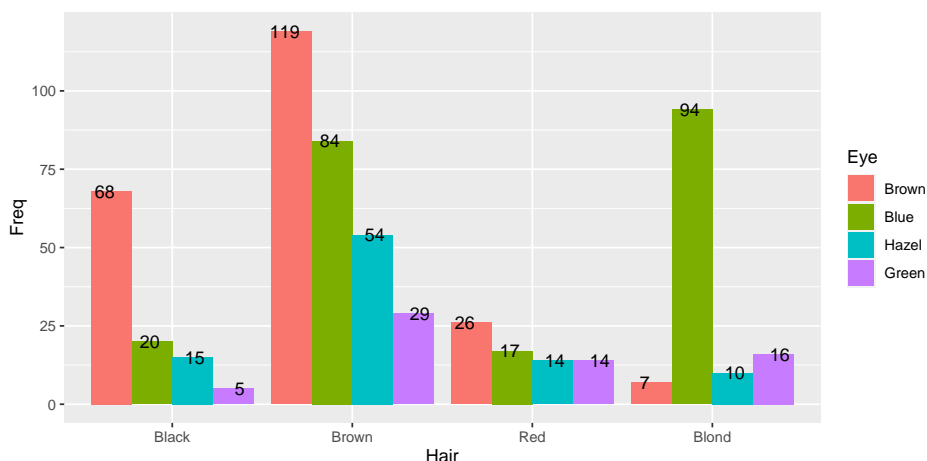
```
df = data.frame(HairEyeColor[,1] + HairEyeColor[,2])
ggplot(df, aes(x=Hair, y=Freq, fill=Eye)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = Freq), position=position_stack(vjust = 0.5))
```



Frequentemente, é necessário fazer comparações da distribuição de frequências de uma variável em vários grupos simultaneamente. Nesse caso, o uso de grá-

ficos bem escolhidos e construídos torna a tarefa muito mais fácil. Na figura abaixo, está representada a distribuição de frequências da cor dos olhos segundo a variável cor do cabelo.

```
df = data.frame(HairEyeColor[, , 1] + HairEyeColor[, , 2])
ggplot(df, aes(x=Hair, y=Freq, fill=Eye)) +
  geom_bar(position="dodge", stat="identity") +
  geom_text(aes(label = Freq), position=position_dodge(width = 1))
```



## 2.3 Variáveis Quantitativas Discretas

Quando estamos trabalhando com uma variável discreta que assume poucos valores, podemos dar a ela o mesmo tratamento dado às variáveis qualitativas ordinais, assumindo que cada valor é uma classe e que existe uma ordem natural nessas classes.

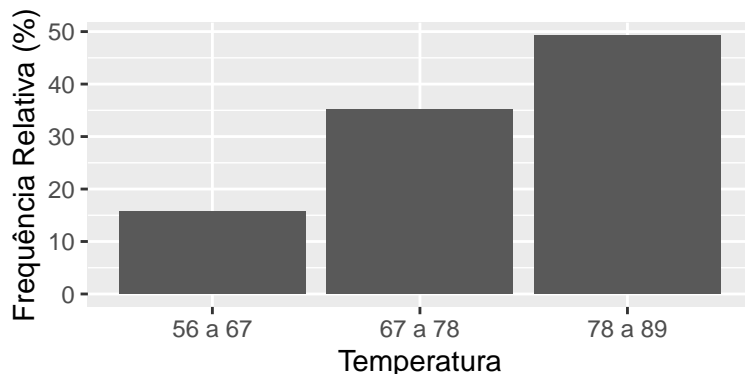
Quando trabalhamos com uma variável discreta que pode assumir um grande número de valores distintos como, por exemplo, parte inteira da temperatura máxima, a construção da tabela de frequências e de gráficos considerando cada valor como uma categoria fica inviável. A solução é agrupar os valores em classes ao montar a tabela, como mostra a tabela abaixo.

Temperatura	Frequência Absoluta	Frequência Relativa (%)	Frequência Absoluta Acumulada	Frequência Relativa Acumulada (%)
56 a 67	21	15.67	21	15.67
67 a 78	47	35.07	68	50.75
78 a 89	66	49.25	134	100
Total	134	99.99	-	-

A Figura abaixo mostra o gráfico da distribuição de frequências da temperatura

medida por 236 dias consecutivos.

```
dt %>% filter(dt$Variavel != "Total") %>%
  ggplot(aes(x=Variavel, y=Freq_rel)) +
  geom_bar(position="dodge",stat="identity") +
  labs(x="Temperatura",y="Frequência Relativa (%)") +
  theme(legend.position = "none")
```



## 2.4 Variáveis Quantitativas Contínuas

Quando a variável em estudo é do tipo contínua, que assume muitos valores distintos, o agrupamento dos dados em classes será sempre necessário na construção das tabelas de frequências. A tabela abaixo apresenta a distribuição de frequências para o peso dos carros.

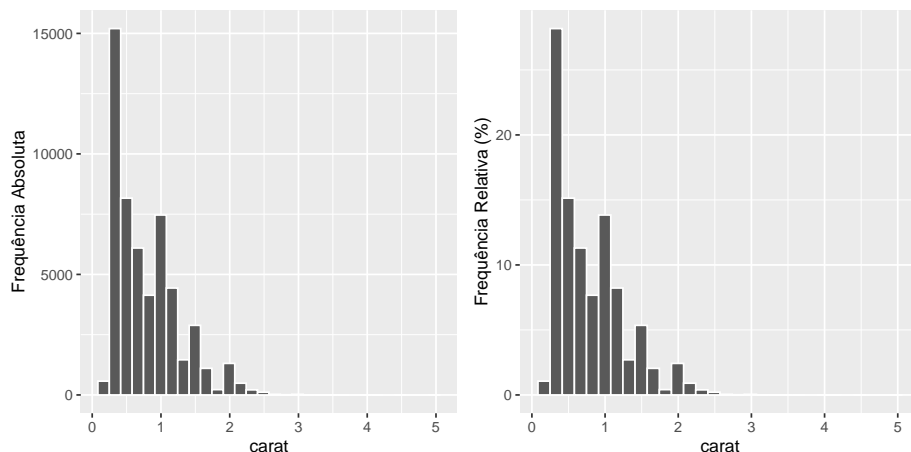
Temperatura	Frequência Absoluta	Frequência Relativa (%)	Frequência Absoluta Acumulada	Frequência Relativa Acumulada (%)
1.5   - 2	0	0.00	0	0
2   - 2.5	0	0.00	0	0
2.5   - 3	3	9.38	3	9.38
3   - 3.5	10	31.25	13	40.62
3.5   - 4	12	37.50	25	78.12
4   - 4.5	6	18.75	31	96.88
4.5   - 5	1	3.12	32	100
5   - 5.5	0	0.00	32	100
Total	32	100.00	-	-

## 2.5 Histograma

A representação gráfica da distribuição de frequências de uma variável contínua é feita através de um gráfico chamado histograma, mostrado nas figuras abaixo

com o peso de diamantes (`diamonds`). O histograma nada mais é do que o gráfico de barras verticais, porém construído com as barras unidas, devido ao caráter contínuo dos valores da variável.

```
library(ggplot2)
p1=ggplot(diamonds,aes(carat)) +
  geom_histogram(color = "white") +
  labs(y = "Frequência Absoluta")
p2=ggplot(diamonds,aes(carat)) +
  geom_histogram(aes(y = (..count..)/sum(..count..)*100),color = "white") +
  labs(y = "Frequência Relativa (%)")
gridExtra::grid.arrange(p1,p2,ncol=2)
```



Os histogramas das figuras acima têm exatamente a mesma forma, apesar de serem construídos usando as frequências absolutas e relativas, respectivamente. O objetivo dessas figuras é mostrar que a escolha do tipo de frequência a ser usada não muda a forma da distribuição. Entretanto, o uso da frequência relativa torna o histograma comparável a outros histogramas, mesmo que os conjuntos de dados tenham tamanhos diferentes (desde a mesma escala seja usada!)

Para o desenvolvimento do histograma:

- Calcule a amplitude dos dados
- Defina a quantidade de classes (as barras verticais). Não existe uma regra, porém uma boa aproximação seria calcular a raiz quadrada da quantidade de dados ( $\sqrt{n}$ )
- Calcule o intervalo das classes dividindo a amplitude pela quantidade de classes
- Determine os limites das classes. Selecione o valor mínimo dos dados (se for mais viável, ele pode ser arredondado para baixo) e soma o valor do intervalo de classe para obter o limite superior
- Faça o passo anterior para todas as classes



- Calcule a frequência dos dados que pertence a cada intervalo

Este é o passo a passo básico para a elaboração de um Histograma, que seja capaz de lhe trazer informações precisas sobre a frequência com que algo acontece em um determinado contexto.

Ao estudarmos a distribuição de frequências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

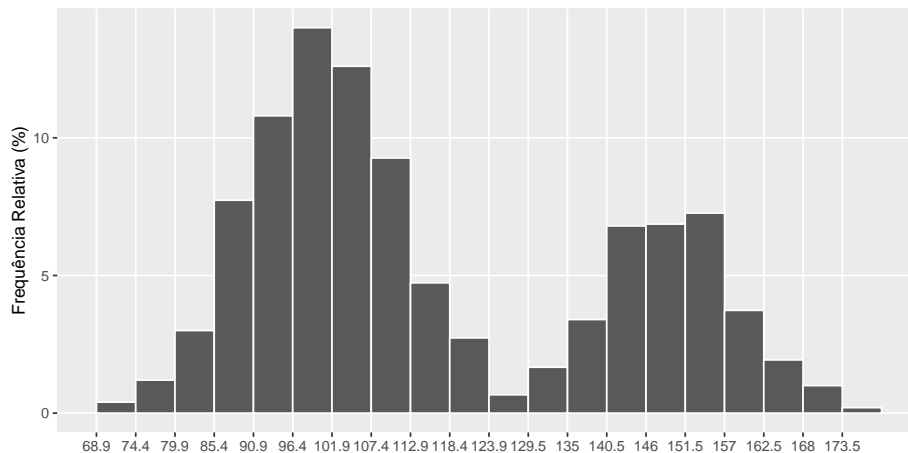
- Tendência Central;
- Variabilidade;
- Forma.

Tais características podem ser quantificadas através das medidas de síntese numérica ou visualizadas a partir do histograma.

### 2.5.1 Tendência Central

A tendência central da distribuição de frequências de uma variável é caracterizada pelo valor (ou faixa de valores) “típico” da variável.

Uma das maneiras de representar o que é “típico” é através do valor mais frequente da variável, chamado de moda. Ou, no caso da tabela de frequências, a classe de maior frequência, chamada de classe modal. No histograma, esta classe corresponde àquela com barra mais alta (“pico”).



No exemplo acima, a classe modal é a que vai de 96.4 até 101.9. Assim, os dados repousam, tipicamente entre esses valores. Entretanto, temos dois picos.

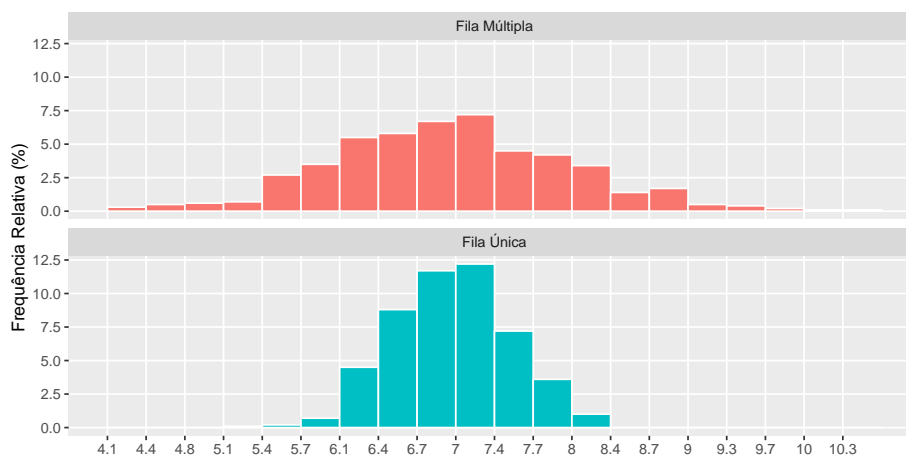
Geralmente, um histograma bimodal indica a existência de dois grupos, com valores centrados em dois pontos diferentes do eixo de valores. Uma distribuição de frequências pode também ser amodal, ou seja, todos os valores são igual-

mente frequentes. Ou também unimodal, quando os valores estão concentrados somente em um ponto/classe.

### 2.5.2 Variabilidade

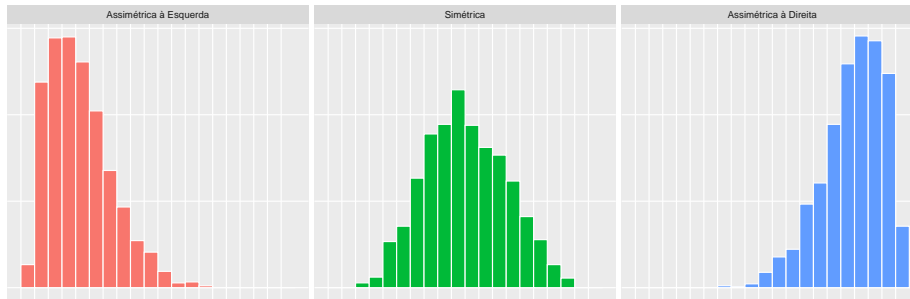
Para descrever adequadamente a distribuição de frequências de uma variável quantitativa, além da informação do valor representativo da variável (tendência central), é necessário dizer também o quanto estes valores variam, ou seja, o quão dispersos eles são.

De fato, somente a informação sobre a tendência central de um conjunto de dados não consegue representá-lo adequadamente. A Figura abaixo mostra um histograma para os tempos de espera de 1000 clientes de dois bancos, um com fila única e outro com fila múltipla, com o mesmo número de atendentes. Os tempos de espera nos dois bancos têm a mesma tendência central de 7 minutos. Entretanto, os dois conjuntos de dados são claramente diferentes, pois os valores são muito mais dispersos no banco com fila múltipla. Assim, quando entramos num fila única, esperamos ser atendidos em cerca de 7 minutos, com uma variação de, no máximo, meio minuto a mais ou a menos. Na fila múltipla, a variação é maior, indicando-se que tanto pode-se esperar muito mais ou muito menos que o valor típico de 7 minutos.



### 2.5.3 Forma

A distribuição de frequências de uma variável pode ter várias formas, mas existem três formas básicas, apresentadas na figura abaixo através de histogramas.



Quando uma distribuição é simétrica em torno de um valor (o mais frequente), significa que as observações estão igualmente distribuídas em torno desse valor (metade acima e metade abaixo).

A assimetria de uma distribuição pode ocorrer de duas formas:

- quando os valores concentram-se à esquerda (assimetria com concentração à esquerda ou assimetria com cauda à direita);
- quando os valores concentram-se à direita (assimetria com concentração à direita ou com assimetria cauda à esquerda);

Ao definir a assimetria de uma distribuição, algumas pessoas preferem se referir ao lado onde está a concentração dos dados. Porém, outras pessoas preferem se referir ao lado onde está “faltando” dados (cauda). As duas denominações são alternativas.

Em alguns casos, apenas o conhecimento da forma da distribuição de frequências de uma variável já nos fornece uma boa informação sobre o comportamento dessa variável. Por exemplo, o que você acharia se soubesse que a distribuição de frequências das notas da primeira prova da disciplina de Estatística que você está cursando é, geralmente, assimétrica com concentração à direita? Como você acha que é a forma da distribuição de frequências da renda no Brasil?

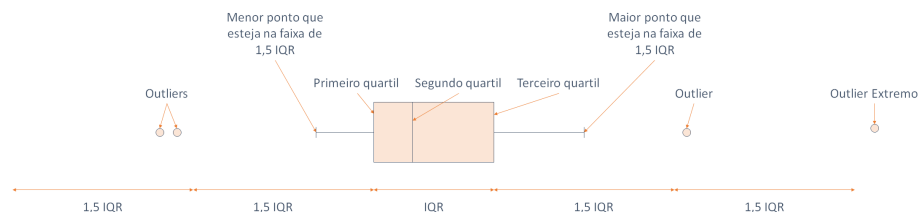
## 2.6 Boxplot

O Boxplot é um gráfico proposto para a detecção de valores discrepantes (outliers), que são aqueles valores muito diferentes do restante do conjunto de dados.

Esses valores discrepantes podem representar erros no processo de coleta ou de processamento dos dados, e, nesse caso, devem ser corrigidos ou excluídos do banco de dados. No entanto, os outliers podem ser valores corretos, que, por alguma razão, são muito diferentes dos demais valores. Nesse caso, a análise desses dados deve ser cuidadosa, pois, como sabemos, algumas estatísticas descritivas, como a média e o desvio-padrão, são influenciadas por valores extremos.

Na construção do Boxplot, utilizamos alguns percentis (mediana, primeiro e terceiro quartis), que são pouco influenciados por valores extremos. Além disso, precisamos saber quais são os valores mínimo e máximo do conjunto de dados.

O Boxplot é constituído por uma caixa atravessada por uma linha, construído usando um eixo com uma escala de valores, como mostra a figura abaixo. Como sabemos, entre o primeiro e o terceiro quartis, temos 50% dos dados. Podemos pensar, então, que essa caixa contém metade dos dados do conjunto.



Como um gráfico tem que representar todos os valores do conjunto de dados, precisamos representar os outros 50%, sendo 25% abaixo do Q1 e 25% acima do Q3. Esses valores serão representados pelas duas linhas que saem das extremidades da caixa. Cada uma das linhas é traçada, a partir das extremidades da caixa, até que encontre o valor máximo ou mínimo; ou atinja o comprimento máximo de 1,5 vezes a altura da caixa (IQR). O que acontecer primeiro.

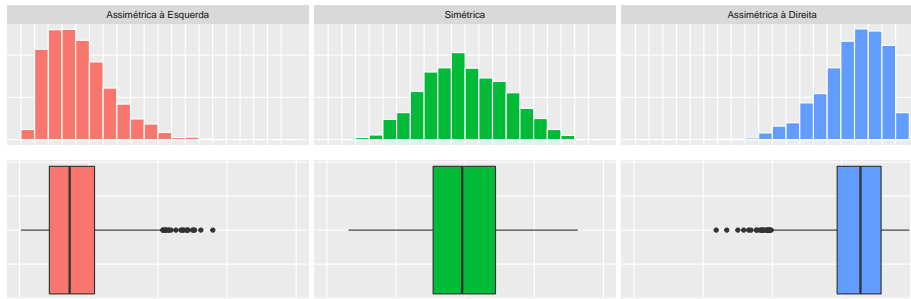
No exemplo da imagem acima o segundo caso aconteceu, assim os valores que ainda não foram representados devem ser devidamente marcados em suas respectivas posições na escala de valores. Esses valores são considerados outliers pelo critério do boxplot. Obviamente, o limite superior do boxplot não coincidiu com o valor máximo do conjunto de dados, que foi considerado um valor discrepante (outlier).

Além da detecção de valores discrepantes, o boxplot pode ser muito útil na análise da distribuição dos valores de um conjunto de dados. Através do boxplot, podemos:

- identificar a forma da distribuição (simétrica ou assimétrica);
- avaliar e comparar a tendência central (mediana) de dois ou mais conjuntos de dados;
- comparar a variabilidade de dois ou mais conjuntos de dados

Para avaliar a forma da distribuição, devemos observar o deslocamento da caixa em relação a linha do boxplot. Lembrando que a caixa do boxplot contém 50% dos dados, o seu deslocamento na linha nos informa onde estão concentrados os dados.

Se a caixa está mais deslocada para um dos lados da linha, significa que metade dos dados estão concentrados naquele lado da escala de valores e, assim, a distribuição é assimétrica. Se a caixa está praticamente no meio da linha, dividindo-a em duas partes iguais, a distribuição será considerada simétrica.

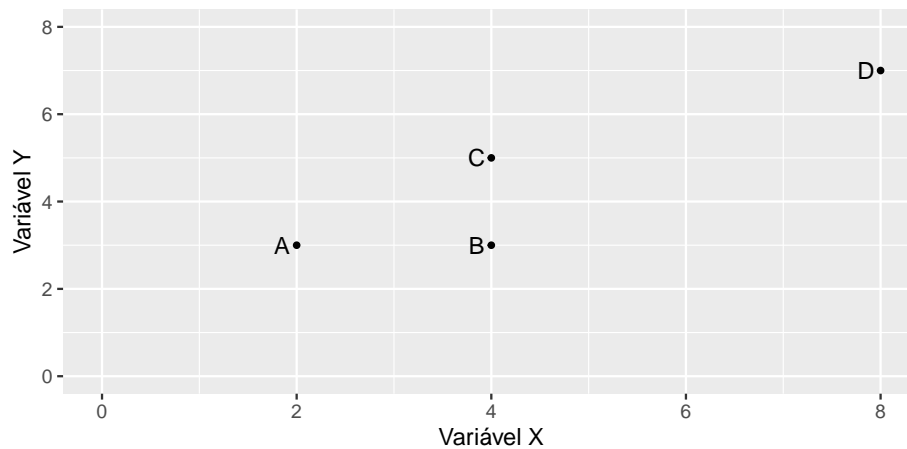


## 2.7 Diagrama de Dispersão

O diagrama de dispersão é um gráfico onde pontos no espaço cartesiano XY são usados para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada indivíduo do conjunto de dados.

O Quadro e a Figura abaixo mostram um esquema do desenho do diagrama de dispersão. Neste exemplo, foram medidos os valores de duas variáveis quantitativas, X e Y, em quatro indivíduos. O eixo horizontal do gráfico representa a variável X e o eixo vertical representa a variável Y.

Indivíduos	Variável X	Variável Y
A	2	3
B	4	3
C	4	5
D	8	7



O diagrama de dispersão é usado principalmente para visualizar a relação/associação entre duas variáveis, mas também para é muito útil para:

- Comparar o efeito de dois tratamentos no mesmo indivíduo

Planta	Quantidade												
1001	5	15.15	15.45	15.63	15.65	16.38	NA	NA	NA	NA	NA	NA	NA
1002	6	14.00	14.50	15.35	15.86	15.94	16.13	NA	NA	NA	NA	NA	NA
1003	7	13.67	13.76	14.06	14.11	14.54	14.89	15.50	NA	NA	NA	NA	NA
1004	8	11.00	11.50	12.39	12.39	12.90	14.50	15.50	16.56	NA	NA	NA	NA
1005	9	10.24	11.12	12.05	12.37	13.48	13.80	14.04	15.39	16.00	NA	NA	NA
1006	10	9.00	9.32	10.67	11.56	11.67	12.56	12.83	12.84	13.43	15.09	NA	NA
1007	11	7.82	8.56	8.74	9.57	11.08	11.92	12.13	12.50	14.14	14.20	14.00	NA
1008	12	7.25	9.41	10.15	10.33	10.80	10.95	11.13	11.48	11.49	12.86	13.37	15.04
1009	13	6.95	7.61	8.53	10.00	10.94	11.04	11.43	11.63	11.97	12.02	12.74	13.53
1010	14	7.00	8.00	9.00	10.00	10.00	10.50	11.00	11.16	11.17	11.70	12.45	12.89

- Verificar o efeito tipo antes/depois de um tratamento

A seguir, veremos dois exemplos da utilização do diagrama de dispersão. O primeiro refere-se ao estudo da associação entre duas variáveis. O segundo utiliza o diagrama de dispersão para comparar o efeito da aplicação de um tratamento, comparando as medidas antes e depois da medicação.

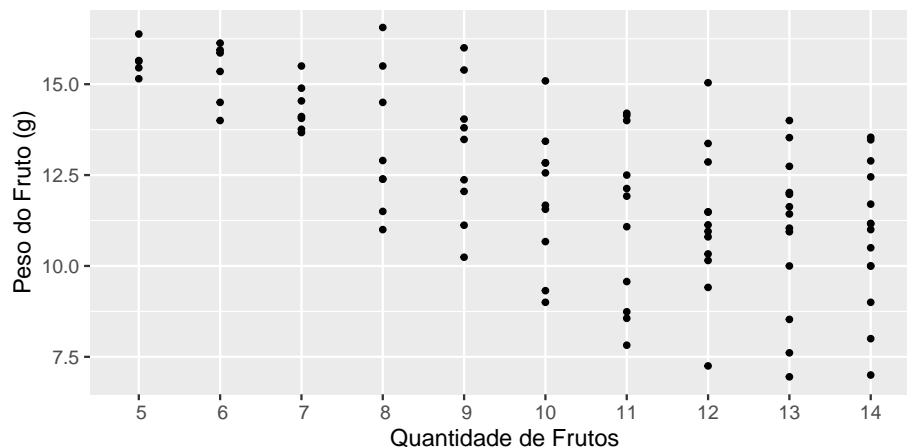
### 2.7.1 Exemplo 1

Um produtor de morangos para exportação deseja produzir frutos grandes, pois frutos pequenos têm pouco valor mesmo no mercado interno. Além disso, os frutos, mesmo grandes, não devem ter tamanhos muito diferentes entre si. O produtor suspeita que uma dos fatores que altera o tamanho dos frutos é o número de frutos por árvore.

Para investigar a relação entre o número de frutos que uma planta produz e o peso destes frutos, ele observou dados de 10 morangueiros na primeira safra (Quadro abaixo) e gerou o Diagrama de Dispersão apresentado abaixo.

```
library(ggplot2)
library(tidyr)
aux = matrix(c(15.15,15.45,15.63,15.65,16.38,NA,NA,NA,NA,NA,NA,NA,NA,NA,
               14,14.5,15.35,15.86,15.94,16.13,NA,NA,NA,NA,NA,NA,NA,NA,NA,
               13.67,13.76,14.06,14.11,14.54,14.89,15.5,NA,NA,NA,NA,NA,NA,NA,
               11,11.5,12.39,12.39,12.9,14.5,15.5,16.56,NA,NA,NA,NA,NA,NA,NA,
               10.24,11.12,12.05,12.37,13.48,13.8,14.04,15.39,16,NA,NA,NA,NA,NA,
               9,9.32,10.67,11.56,11.67,12.56,12.83,12.84,13.43,15.09,NA,NA,NA,NA,
               7.82,8.56,8.74,9.57,11.08,11.92,12.13,12.5,14.14,14.2,14,NA,NA,NA,
               7.25,9.41,10.15,10.33,10.8,10.95,11.13,11.48,11.49,12.86,13.37,15.04,NA,
               6.95,7.61,8.53,10,10.94,11.04,11.43,11.63,11.97,12.02,12.74,13.53,14,NA,
               7,8,9,10,10,10.5,11,11.16,11.17,11.7,12.45,12.89,13.47,13.54),
             nrow=10,ncol = 14,byrow = T)
dados = data.frame(ID_Morangueiro = 1001:1010,
                   Qtd_Frutos = 5:14,aux)
dados %>%
  pivot_longer(starts_with("X")) %>%
  ggplot(aes(x = Qtd_Frutos, y = value)) +
  geom_point(size = 1) +
```

```
scale_x_discrete(limits=5:14) +  
labs(x = "Quantidade de Frutos", y = "Peso do Fruto (g)")
```



O diagrama de dispersão mostra-nos dois fatos. O primeiro, que há um decréscimo no valor médio do peso do fruto por árvore à medida que cresce o número de frutos na árvore. Ou seja, não é vantagem uma árvore produzir muitos frutos, pois eles tenderão a ser muito pequenos.

O segundo fato que percebemos é que, com o aumento no número de frutos na árvore, cresce também a variabilidade no peso, gerando tanto frutos muito grandes, como muito pequenos.

Assim, conclui-se que não é vantagem ter poucas plantas produzindo muito frutos, mas sim muitas plantas produzindo poucos frutos, mas grandes e uniformes. Uma análise mais detalhada poderá determinar o número ideal de frutos por árvore, aquele que maximiza o peso médio e, ao mesmo tempo, minimiza a variabilidade do peso.

### 2.7.2 Exemplo 2

Captopril é um remédio destinado a baixar a pressão sistólica. Para testar seu efeito, ele foi ministrado a 12 pacientes, tendo sido medida a pressão sistólica antes e depois da medicação.

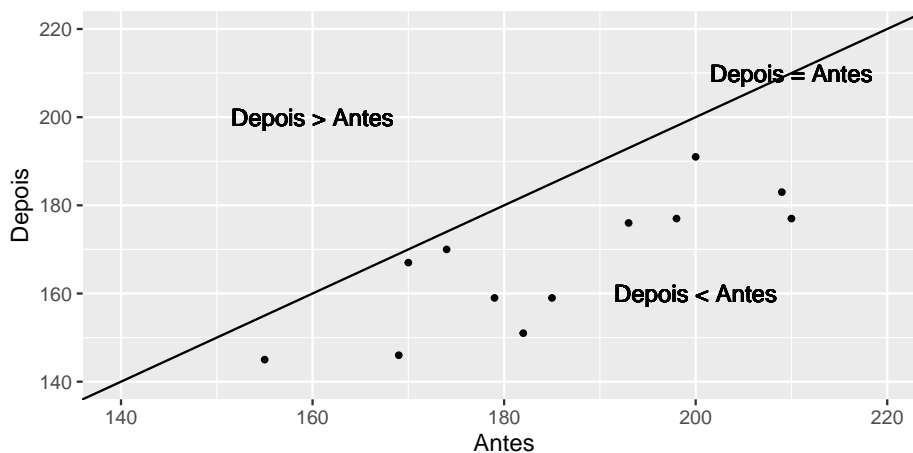
Paciente	Antes	Depois
A	200	191
B	174	170
C	198	177
D	170	167
E	179	159
F	182	151
G	193	176
H	209	183
I	185	159
J	155	145
K	169	146
L	210	177

Os mesmos indivíduos foram utilizados nas duas amostras (Antes/depois). Assim, é natural compararmos a pressão sistólica para cada indivíduo, comparando a pressão sistólica depois e antes. Para todos os pacientes, a pressão sistólica depois do Captopril é menor do que antes da medicação. Mas como podemos “ver” se estas diferenças são grandes? Com a ajuda do diagrama de dispersão mostrado na figura abaixo.

```
dados = data.frame(Paciente = LETTERS[1:12],
                   Antes = c(200, 174, 198, 170, 179, 182, 193, 209, 185, 155, 169, 210),
                   Depois = c(191, 170, 177, 167, 159, 151, 176, 183, 159, 145, 146, 177))

dados %>%
  ggplot(aes(x = Antes, y = Depois)) +
  geom_point(size = 1) +
  ylim(140,220) + xlim(140,220) +
  geom_abline(slope = 1, intercept = 0) +
  geom_text(aes(x = 200, y = 160), label = "Depois < Antes") +
  geom_text(aes(x = 160, y = 200), label = "Depois > Antes") +
  geom_text(aes(x = 210, y = 210), label = "Depois = Antes")
```





Cada ponto no diagrama de dispersão corresponde às medidas de pressão sistólica de um paciente, medida antes e depois da medicação. A linha marcada no diagrama corresponde à situação onde a pressão sistólica não se alterou depois do paciente tomar o Captopril. Veja que todos os pontos estão abaixo desta linha, ou seja para todos os pacientes o Captopril fez efeito. Grande parte destes pontos está bem distante da linha, mostrando que a redução na pressão sistólica depois do uso do medicamento não foi pequena.

## 2.8 Séries Temporais

Séries temporais (ou séries históricas) são um conjunto de observações de uma mesma variável quantitativa (discreta ou contínua) feitas ao longo do tempo. O conjunto de novos casos da COVID-19 é um exemplo de série temporal.

Um dos objetivos do estudo de séries temporais é conhecer o comportamento da série ao longo do tempo (aumento, estabilidade ou declínio dos valores). Em alguns estudos, esse conhecimento pode ser usado para se fazer previsões de valores futuros com base no comportamento dos valores passados.

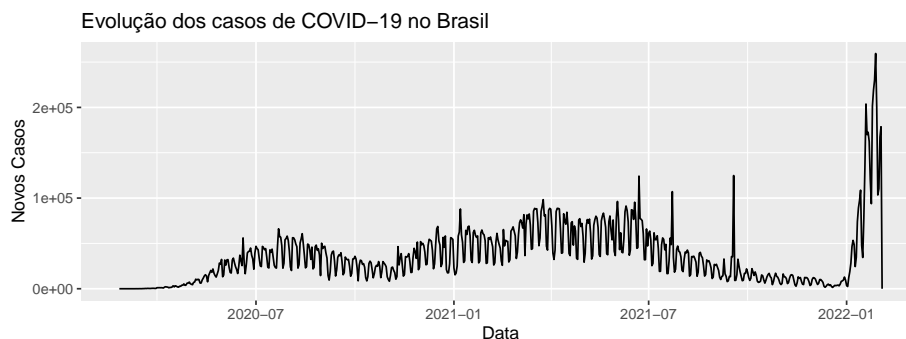
A representação gráfica de uma série temporal é feita através do gráfico de linha, como exemplificado nas figuras abaixo. No eixo horizontal do gráfico de linha, está o indicador de tempo e, no eixo vertical, a variável a ser representada.

```
library(tsibble)
library(lubridate)
library(dplyr)

dados = read.csv("https://raw.githubusercontent.com/wcota/covid19br/master/cases-brazil-states.csv")

dados %>%
  mutate(date = as_date(date)) %>%
  as_tsibble(key = state, index = date) %>%
```

```
filter(state == "TOTAL") %>%
select(date, newCases) %>%
ggplot(aes(x = date, y = newCases)) +
geom_line() +
labs(title = "Evolução dos casos de COVID-19 no Brasil",
      y = "Novos Casos", x = "Data")
```



De maneira geral, um gráfico de linhas deve ser construído de modo que:

- O início do eixo vertical seja o valor mínimo possível para a variável que está sendo representada, para evitar as distorções
- O final do eixo vertical seja tal que a série fica centrada em relação ao eixo vertical
- Os tamanhos dos eixos sejam o mais parecidos possível

## 2.9 Gráfico de Radar

O gráfico de radar é um diagrama e/ou gráfico que consiste de uma sequência de raios equi-angulares, com cada raio representando uma das variáveis. O comprimento de cada raio é proporcional à magnitude da variável para o ponto de dados em relação à máxima magnitude da variável em todos os pontos. Uma linha é desenhada ligando os valores de cada raio. Isso dá ao diagrama uma aparência de estrela, o que deu origem a um dos nomes mais populares para este gráfico. O gráfico de estrela pode ser usado para responder as seguintes questões:

- Que observações são mais semelhantes, por exemplo, existem clusters de observações?
- Existem exceções?

Gráficos de radar oferecem uma maneira útil de exibir observações multivariáveis com um número arbitrário de variáveis. Cada estrela representa uma única observação. Normalmente, os gráficos de radar são gerados em um formato multi-diagrama com muitas estrelas em cada página, cada estrela repre-

sentando uma observação. É um pouco mais fácil de ver padrões em dados se as observações forem organizadas em alguma ordem não-arbitrária (se as variáveis forem atribuídas aos raios da estrela em uma ordem significativa).

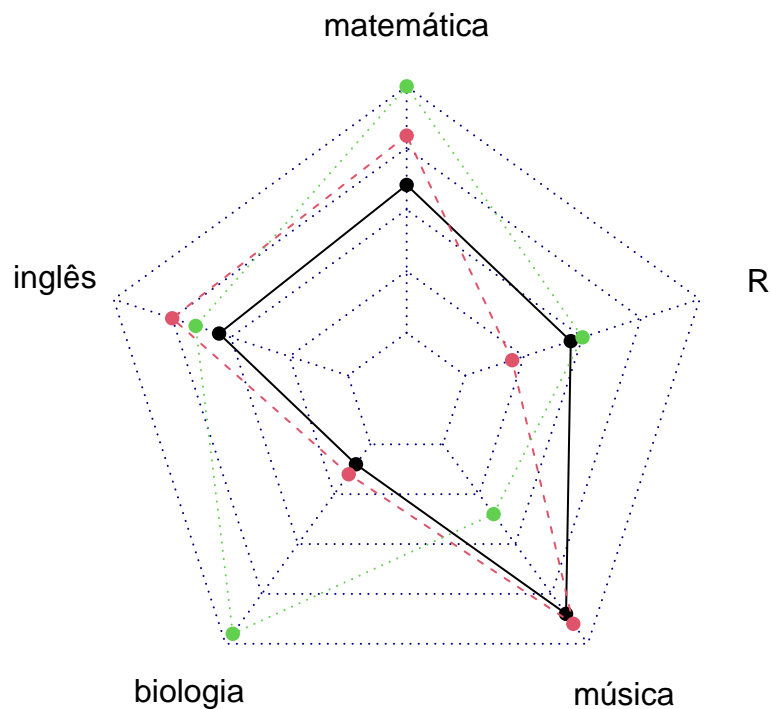
Uma aplicação de gráficos de radar é o controle de melhoria de qualidade para apresentar as métricas de desempenho de qualquer programa em curso. Eles também são usados em esportes para representar os pontos fortes e fracos de jogadores, onde eles são geralmente chamados de gráficos de aranha. No exemplo abaixo vemos a performance de três alunos (a,b,c) em relação as dimensões matemática, inglês, biologia, música e R.

```
library(fmsb)

dados = as.data.frame(matrix(sample(0:20, 15, replace=F), ncol=5))
colnames(dados) = c("matemática", "inglês", "biologia", "música", "R")
rownames(dados) = paste("aluno", letters[1:3])

dados = rbind(rep(20,5), rep(0,5), dados)

radarchart(dados)
```





# Bibliography

Wesley Cota. Monitoring the number of COVID-19 cases and deaths in brazil at municipal and federative units level. *SciELOPreprints:362*, May 2020. doi: 10.1590/scielopreprints.362. URL <https://doi.org/10.1590/scielopreprints.362>.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>.