

Probabilidade e Estatística com R

Fernando Náufel

(versão de 04/05/2022)

Sumário

Apresentação	3
Referências recomendadas	5
Exercício	5
1 O Que É Estatística?	6
1.1 Vídeo 1	6
1.2 Exercícios	6
1.3 Vídeo 2	7
1.4 Exercícios	9
2 Introdução a R	10
2.1 Vídeo 1	10
2.2 Vídeo 2	10
2.3 Exercícios	10
3 Introdução ao tidyverse	12
3.1 Criando uma <i>tibble</i>	12
3.2 Operador de <i>pipe</i> (<code>%>%</code>)	14
3.3 Formato <i>tidy</i>	15
3.4 Manipulando os dados	20
3.5 Exercícios	34
4 Visualização com ggplot2	37
4.1 Vídeo 1	37
4.2 Componentes de um gráfico ggplot2	37
4.3 Conjunto de dados	40
4.4 Gráficos de dispersão (<i>scatter plots</i>)	46
4.5 Vídeo 2	63
4.6 Histogramas e cia.	63
4.7 Ogiva	68
4.8 Ramos e folhas	69

4.9	Personalização do tema	70
4.10	Exercícios	73
5	Visualização com ggplot2 (continuação)	77
5.1	Vídeo 1	77
5.2	<i>Boxplots</i>	77
5.3	Vídeo 2	87
5.4	Gráficos de barras e de colunas	87
5.5	Gráficos de linha e séries temporais	101
5.6	Exercícios	106
5.7	Referências sobre visualização e R	106
6	Medidas	107
6.1	Vídeo	107
6.2	Medidas de centralidade	107
6.3	Formas de uma distribuição	115
6.4	Re-expressão	121
6.5	Medidas de posição	122
6.6	Medidas de dispersão	123
6.7	Coeficiente de variação	130
6.8	Escores-padrão	132
6.9	Teorema de Tchebychev	134
7	Probabilidades	138
7.1	Vídeo 1	138
7.2	Espaço amostral	138
7.3	Evento	139
7.4	Análise Combinatória	139
7.5	Probabilidade clássica	140
7.6	Probabilidade empírica	140
7.7	Probabilidade subjetiva	143
7.8	Formalização de probabilidades	144
7.9	Eventos independentes (explicação informal)	144
7.10	$P(A \cup B)$ com A e B não-disjuntos	145
7.11	Problema do aniversário	146
7.12	Exercícios	153
7.13	Vídeo 2	154
7.14	Probabilidade condicional	154
7.15	Probabilidade conjunta	160
7.16	Independência	161
7.17	Probabilidade total	163
7.18	Teorema de Bayes	164
	Referências	168

Apresentação

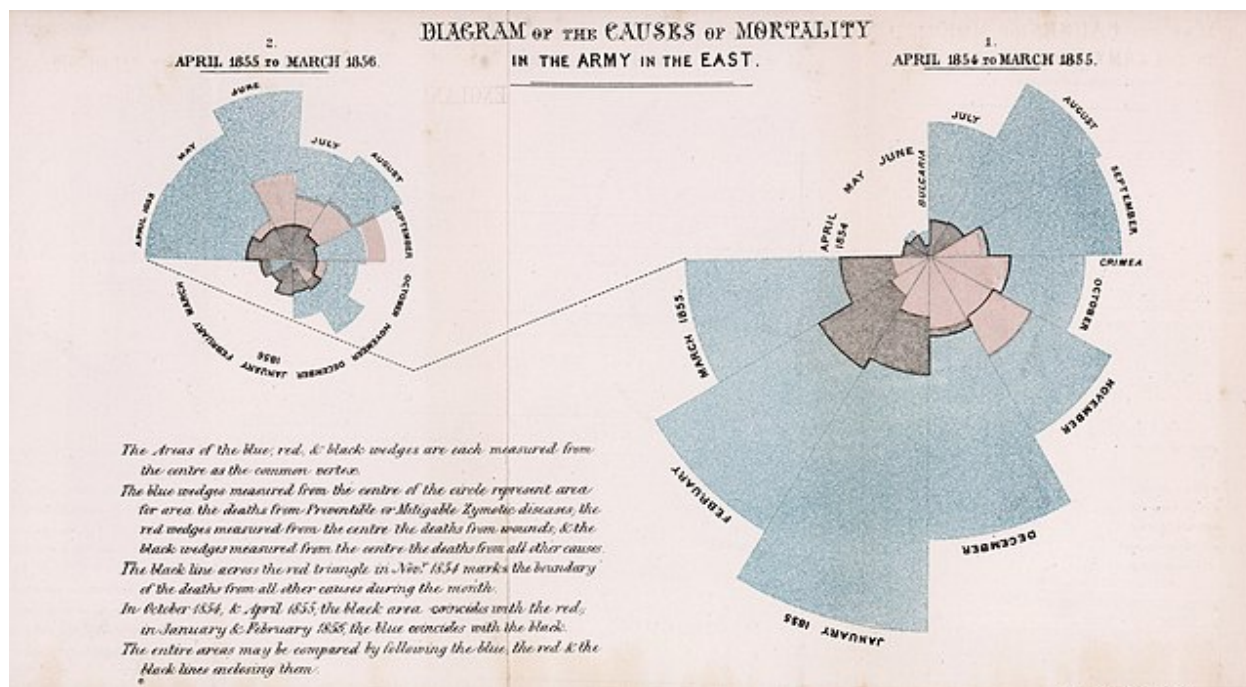
Atenção



Este material ainda está em construção.

Pode haver mudanças a qualquer momento.

Verifique, no rodapé da página *web* ou na capa do arquivo pdf, a data desta versão.



Este livro/site foi iniciado em 2020, durante a pandemia de COVID-19, quando a Universidade Federal Fluminense (UFF) funcionou em regime de ensino remoto durante mais de um ano.

Para atender os alunos do curso de Probabilidade e Estatística do curso de graduação em Ciência da Computação da UFF, decidi gravar aulas em vídeo e disponibilizar os arquivos usados nelas. Foram esses arquivos que deram origem a este livro/site.

Este livro/site foi construído para pessoas que já saibam programar, embora não necessariamente em R.

Para tirar o máximo proveito deste material, você deve fazer o seguinte:

1. Assistir aos vídeos contidos em cada capítulo. A *playlist* completa está em <https://www.youtube.com/playlist?list=PL7SRLwLs7ocaV-Y1vrVU3W7mZnnS0qkWW>.
2. Instalar o R no seu computador ou abrir uma conta no RStudio Cloud, para poder usar o R *online*. Você encontra instruções para fazer isto no [capítulo de introdução a R](#).
3. Baixar, [neste repositório do Github](#), o código-fonte deste livro/site, para poder rodar e alterar os exemplos.
4. Seguir os *links* para outras fontes *online* que abordam assuntos que não são cobertos em detalhes neste curso.
5. Fazer os exercícios. Ao longo do tempo, acrescentarei *links* para vídeos explicando as soluções.



Se você estiver lendo este material na web, você pode clicar nos comandos e funções que aparecem nos blocos de código em R para abrir páginas da documentação sobre eles.

Se você preferir ler este livro em pdf, ou se quiser imprimi-lo, [faça o download do arquivo aqui](#).

Referências recomendadas

Em português

- Sillas Gonzaga, *Introdução a R para Visualização e Apresentação de Dados*, http://sillasgonzaga.com/material/curso_visualizacao/index.html
- Allan Vieira de Castro Quadros, *Introdução à Análise de Dados em R utilizando Tidyverse*, https://allanvc.github.io/book_IADR-T/
- Paulo Felipe de Oliveira, Saulo Guerra, Robert McDonnel, *Ciência de Dados com R – Introdução*, <https://cdr.ibpad.com.br/index.html>
- Curso R, *Ciência de Dados em R*, <https://livro.curso-r.com/>

Em inglês

- Garrett Grolemund, Hadley Wickham, *R for Data Science*, <https://r4ds.had.co.nz/>
- Chester Ismay, Albert Y. Kim, *A ModernDive into R and the Tidyverse*, <https://moderndive.com/>

Exercício

1. Pesquise sobre a imagem do início deste capítulo. Ela foi criada em 1858 por Florence Nightingale.

O Que É Estatística?

1.1

Vídeo 1

https://youtu.be/6Q_XSoLCIpc

1.2

Exercícios

1. Você está interessado em estimar a altura de todos os homens da sua faculdade. Para isso, você decide medir as alturas de todos os homens da sua turma de Estatística.
 - Qual é a amostra?
 - Qual é a população?
2. Um instituto de pesquisa entrevista um grupo de 1000 pessoas, perguntando a cada uma se ela vai votar a favor do candidato A na próxima eleição. Dos entrevistados, 600 responderam que sim. A proporção 0,6 (ou 60%) é uma estatística ou um parâmetro?
3. Você vê alguma diferença entre as cinco situações abaixo? Quais das situações são equivalentes em termos da probabilidade de conseguir 10 cartas do mesmo naipe?
 - a. Usando um baralho normal, você retira 10 cartas e registra as cartas retiradas.

- b. Usando um baralho normal, você repete a seguinte sequência de ações 10 vezes: retirar uma carta do baralho, registrar a carta retirada e repor a carta no baralho.
 - c. Usando uma caixa contendo todas as cartas de 1 milhão de baralhos reunidos, você retira 10 cartas e registra as cartas retiradas.
 - d. Usando uma caixa contendo todas as cartas de 1 milhão de baralhos reunidos, você repete a seguinte sequência de ações 10 vezes: retirar uma carta da caixa, registrar a carta retirada e repor a carta na caixa.
 - e. Usando um baralho *infinito*, você retira 10 cartas e registra as cartas retiradas.
 - f. Usando um baralho *infinito*, você repete a seguinte sequência de ações 10 vezes: retirar uma carta do baralho, registrar a carta retirada e repor a carta no baralho.
4. Qual a graça dos quadrinhos na Figura 1.1, que também [aparecem no vídeo](#)?



Figura 1.1: <http://xkcd.com/552/>

5. Qual a graça dos quadrinhos na Figura 1.2?
6. Veja este vídeo sobre o cavalo Hans:

<https://youtu.be/G3VkCmdUfZE>

Qual a relação entre esta história e a necessidade de duplo cegamento?

1.3

Vídeo 2

<https://youtu.be/492VASxIDRo>



Figura 1.2: <http://xkcd.com/1462/>

1.4

Exercícios

1. Por que não faz sentido calcular a média dos CEPs de um grupo de pessoas?
2. Uma temperatura de -40 graus Celsius é igual a uma temperatura de -40 graus Fahrenheit?
3. Uma temperatura de zero graus Celsius é igual a uma temperatura de zero graus Fahrenheit?
4. Uma variação de temperatura de 1 grau Celsius é igual a uma variação de temperatura de 1 grau Fahrenheit?
5. Um saldo bancário de zero reais é igual a um saldo bancário de zero dólares?
6. Um produto de 1 milhão de reais custa o mesmo que um produto de 1 milhão de dólares?
7. Meses representados por números de 1 a 12 são dados de que nível?

Introdução a R

2.1

Vídeo 1

<https://youtu.be/1kXQDNqm41c>

2.2

Vídeo 2

<https://youtu.be/3GEc1oiKDrU>

2.3

Exercícios

1. Para criar sua conta no RStudio Cloud, acesse <https://rstudio.cloud/>.
2. Se você preferir instalar o R no seu computador, acesse
 - <https://cran.r-project.org/> para baixar e instalar o R, e
 - <https://rstudio.com/products/rstudio/download/> para baixar e instalar o RStudio, um IDE específico para R.

3. Abra o RStudio Cloud ou o seu RStudio instalado localmente.
4. Crie um novo projeto. Sempre trabalhe em projetos para ter seus arquivos organizados.
5. Para instalar o `swirl` (pacote do R para exercícios interativos), execute o seguinte comando no console do RStudio:

```
install.packages("swirl")
```

6. Para instalar os exercícios de introdução a R, execute os seguintes comandos no console do RStudio:

```
library(swirl)
install_course_github('fnaufel', 'introR')
```

7. Mude o idioma para português e execute o `swirl`.

```
select_language('portuguese', append_rprofile = TRUE)
swirl()
```

8. Na primeira execução, você vai precisar se identificar (qualquer nome serve). Com essa identificação, o `swirl` vai registrar o seu progresso nas lições.
9. No `swirl`, as perguntas são mostradas no console. Você também deve responder no console.
10. Às vezes, um *script* será aberto no editor de textos para que você complete um programa. Quando seu programa estiver pronto, salve o arquivo e digite `submit()` no console para o `swirl` processar o *script*.
11. O `swirl` dá instruções claras no console. Na dúvida, digite `info()` no *prompt* do R (`>`).
12. Se, em vez do *prompt* do R, o console mostrar reticências (`...`), tecle *Enter*.
13. Se nada funcionar, tecle *ESC*.
14. Para sair do `swirl()`, digite `bye()` no *prompt* do R.
15. Para voltar para os exercícios, digite

```
library(swirl)
swirl()
```

Introdução ao tidyverse



Busque mais informações sobre os pacotes que compõem o tidyverse [nas referências recomendadas](#).

3.1

Criando uma *tibble*

- Uma *tibble* é uma tabela retangular.
- Cada coluna é um vetor:

```
cores <- tibble(  
  pessoa = c('João', 'Maria', 'Pedro', 'Ana'),  
  'cor favorita' = c('azul', 'rosa', 'preto', 'branco')  
)
```

```
cores
```

```
## # A tibble: 4 x 2  
##   pessoa `cor favorita`  
##   <chr>   <chr>  
## 1 João   azul  
## 2 Maria  rosa
```

```
## 3 Pedro  preto
## 4 Ana    branco
```

- Isto é um pouco diferente da maneira como estamos acostumados a ver tabelas (como uma coleção de linhas, em vez de uma coleção de colunas).
- A função `tribble` permite a entrada de forma mais natural, linha a linha. **Lembre-se de usar `~` antes dos nomes das colunas.**

```
cores <- tribble(
  ~pessoa, ~'cor favorita',
  "João",   "azul",
  "Maria",  "rosa",
  "Pedro",  "preto",
  "Ana",    "branco"
)

cores
```

```
## # A tibble: 4 x 2
##   pessoa `cor favorita`
##   <chr>   <chr>
## 1 João   azul
## 2 Maria  rosa
## 3 Pedro  preto
## 4 Ana    branco
```



Mesmo que você crie uma *tibble* linha a linha, o R vai continuar tratando sua *tibble* como uma coleção de colunas.

É importante lembrar disto para entender a forma como R manipula estas tabelas.

- Se uma coluna não puder ser armazenada em um vetor, a coluna será uma lista (com vetores como elementos):

```
cores <- tibble(
  pessoa = c('João', 'Maria', 'Pedro', 'Ana'),
  'cor favorita' = list(
    c('azul', 'roxo'),
    c('rosa', 'magenta'),
    NA,
    'branco'
  )
)

cores
```

```
## # A tibble: 4 x 2
```

```
##  pessoa `cor favorita`
##  <chr>  <list>
## 1 João   <chr [2]>
## 2 Maria  <chr [2]>
## 3 Pedro  <lgl [1]>
## 4 Ana    <chr [1]>
```

- Use `View()` para examinar interativamente o conteúdo de uma coluna-lista:

```
cores %>% View()
```

3.2

Operador de *pipe* (`%>%`)

- O tidyverse inclui o pacote `magrittr`, que contém o operador `%>%`, chamado *pipe*.¹
- A ideia é facilitar a leitura de **composições de funções**. O código

```
y <- h(g(f(x)))
```

pode ser escrito como

```
y <- x %>% f() %>% g() %>% h()
```

- Esta segunda versão é mais fiel à ordem em que as operações acontecem.
- Na verdade, R tem um operador de **atribuição para a direita**, mas poucas pessoas recomendam usá-lo:

```
x %>% f() %>% g() %>% h() -> y
```

- Se `f`, `g` e `h` forem funções de um argumento só, os parênteses podem ser omitidos:

```
y <- x %>% f %>% g %>% h
```

- Se a função `f` tiver outros argumentos, escreva-os normalmente na chamada a `f`:

```
y <- x %>% mean(na.rm = TRUE)
```

- O *pipe* `EXP %>% f(...)` sempre insere o resultado da expressão `EXP` como o **primeiro argumento da função `f`**.
- Se você precisar que o resultado da expressão `EXP` seja inserido em outra posição na lista de argumentos de `f`, use um ponto “.” para isso:

¹Por que o nome do pacote e o nome do operador formam um trocadilho?

```
x %>% consultar(df, .)
```

equivale a

```
consultar(df, x)
```

3.3

Formato *tidy*

- Nossa última versão da *tibble* `cores` é um pouco mais complexa do que deveria ser:

```
cores
```

```
## # A tibble: 4 x 2
##   pessoa `cor favorita`
##   <chr>   <list>
## 1 João   <chr [2]>
## 2 Maria  <chr [2]>
## 3 Pedro  <lgl [1]>
## 4 Ana    <chr [1]>
```

- O formato *tidy* exige que
 1. Cada linha da *tibble* corresponda a uma observação sobre um indivíduo,
 2. Cada coluna corresponda a uma variável observada, e
 3. Cada célula contenha um valor da variável.
- Na *tibble* `cores`, a primeira e a segunda exigências são satisfeitas, mas a terceira não, pois algumas células contêm valores múltiplos.
- A *tibble* não está no formato *tidy*.
- Podemos “extrair” estes vetores “aninhados” usando o comando `unnest`, do pacote `tidyr`:

```
cores <- cores %>%
  unnest(`cor favorita`)

cores
```

```
## # A tibble: 6 x 2
##   pessoa `cor favorita`
##   <chr>   <chr>
## 1 João   azul
## 2 João   roxo
## 3 Maria  rosa
## 4 Maria  magenta
```



```
## 5 Pedro <NA>
## 6 Ana branco
```

- A maioria das funções do `tidyverse` exige que as *tibbles* estejam neste formato *tidy*.
- Um exemplo mais complexo é o *dataset* `billboard`, com as seguintes colunas (para cada música que estava no *top 100* da Billboard no ano de 2000):
 - Nome do artista ou banda;
 - Nome da música;
 - Data em que a música entrou no *top 100* da Billboard;
 - Para cada uma das 76 semanas seguintes, a posição da música no *top 100*.

```
billboard %>% glimpse()
```

```
## Rows: 317
## Columns: 79
## $ artist      <chr> "2 Pac", "2Ge+her", "3 Doors Down", "3 Doors Dow~
## $ track       <chr> "Baby Don't Cry (Keep...", "The Hardest Part Of ~
## $ date.entered <date> 2000-02-26, 2000-09-02, 2000-04-08, 2000-10-21,~
## $ wk1         <dbl> 87, 91, 81, 76, 57, 51, 97, 84, 59, 76, 84, 57, ~
## $ wk2         <dbl> 82, 87, 70, 76, 34, 39, 97, 62, 53, 76, 84, 47, ~
## $ wk3         <dbl> 72, 92, 68, 72, 25, 34, 96, 51, 38, 74, 75, 45, ~
## $ wk4         <dbl> 77, NA, 67, 69, 17, 26, 95, 41, 28, 69, 73, 29, ~
## $ wk5         <dbl> 87, NA, 66, 67, 17, 26, 100, 38, 21, 68, 73, 23,~
## $ wk6         <dbl> 94, NA, 57, 65, 31, 19, NA, 35, 18, 67, 69, 18, ~
## $ wk7         <dbl> 99, NA, 54, 55, 36, 2, NA, 35, 16, 61, 68, 11, 2~
## $ wk8         <dbl> NA, NA, 53, 59, 49, 2, NA, 38, 14, 58, 65, 9, 17~
## $ wk9         <dbl> NA, NA, 51, 62, 53, 3, NA, 38, 12, 57, 73, 9, 17~
## $ wk10        <dbl> NA, NA, 51, 61, 57, 6, NA, 36, 10, 59, 83, 11, 1~
## $ wk11        <dbl> NA, NA, 51, 61, 64, 7, NA, 37, 9, 66, 92, 1, 17,~
## $ wk12        <dbl> NA, NA, 51, 59, 70, 22, NA, 37, 8, 68, NA, 1, 3,~
## $ wk13        <dbl> NA, NA, 47, 61, 75, 29, NA, 38, 6, 61, NA, 1, 3,~
## $ wk14        <dbl> NA, NA, 44, 66, 76, 36, NA, 49, 1, 67, NA, 1, 7,~
## $ wk15        <dbl> NA, NA, 38, 72, 78, 47, NA, 61, 2, 59, NA, 4, 10~
## $ wk16        <dbl> NA, NA, 28, 76, 85, 67, NA, 63, 2, 63, NA, 8, 17~
## $ wk17        <dbl> NA, NA, 22, 75, 92, 66, NA, 62, 2, 67, NA, 12, 2~
## $ wk18        <dbl> NA, NA, 18, 67, 96, 84, NA, 67, 2, 71, NA, 22, 2~
## $ wk19        <dbl> NA, NA, 18, 73, NA, 93, NA, 83, 3, 79, NA, 23, 2~
## $ wk20        <dbl> NA, NA, 14, 70, NA, 94, NA, 86, 4, 89, NA, 43, 4~
## $ wk21        <dbl> NA, NA, 12, NA, NA, NA, NA, NA, 5, NA, NA, 44, 4~
## $ wk22        <dbl> NA, NA, 7, NA, NA, NA, NA, NA, 5, NA, NA, NA, 50~
## $ wk23        <dbl> NA, NA, 6, NA, NA, NA, NA, NA, 6, NA, NA, NA, NA~
## $ wk24        <dbl> NA, NA, 6, NA, NA, NA, NA, NA, 9, NA, NA, NA, NA~
## $ wk25        <dbl> NA, NA, 6, NA, NA, NA, NA, NA, 13, NA, NA, NA, N~
## $ wk26        <dbl> NA, NA, 5, NA, NA, NA, NA, NA, 14, NA, NA, NA, N~
## $ wk27        <dbl> NA, NA, 5, NA, NA, NA, NA, NA, 16, NA, NA, NA, N~
```

[illegible]

```
## $ wk76          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

- Vamos renomear as colunas:

```
bb <- billboard %>%
  rename(
    artista = artist,
    musica = track,
    entrou = date.entered
  )
```

```
bb %>% head()
```

```
## # A tibble: 6 x 79
##   artista  musica entrou      wk1    wk2    wk3    wk4    wk5    wk6    wk7
##   <chr>    <chr> <date>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2 Pac    Baby ~ 2000-02-26    87    82    72    77    87    94    99
## 2 2Ge+her  The H~ 2000-09-02    91    87    92    NA    NA    NA    NA
## 3 3 Doors ~ Krypt~ 2000-04-08    81    70    68    67    66    57    54
## 4 3 Doors ~ Loser 2000-10-21    76    76    72    69    67    65    55
## 5 504 Boyz Wobbl~ 2000-04-15    57    34    25    17    17    31    36
## 6 98^0     Give ~ 2000-08-19    51    39    34    26    26    19     2
## # ... with 69 more variables: wk8 <dbl>, wk9 <dbl>, wk10 <dbl>,
## #   wk11 <dbl>, wk12 <dbl>, wk13 <dbl>, wk14 <dbl>, wk15 <dbl>,
## #   wk16 <dbl>, wk17 <dbl>, wk18 <dbl>, wk19 <dbl>, wk20 <dbl>,
## #   wk21 <dbl>, wk22 <dbl>, wk23 <dbl>, wk24 <dbl>, wk25 <dbl>,
## #   wk26 <dbl>, wk27 <dbl>, wk28 <dbl>, wk29 <dbl>, wk30 <dbl>,
## #   wk31 <dbl>, wk32 <dbl>, wk33 <dbl>, wk34 <dbl>, wk35 <dbl>,
## #   wk36 <dbl>, wk37 <dbl>, wk38 <dbl>, wk39 <dbl>, wk40 <dbl>, ...
```

- O que é uma observação neste conjunto de dados?
A posição, em uma semana, de uma música que esteve no *top 100* da *Billboard* durante o ano 2000.
- Quais são as variáveis que qualificam cada observação?
 - O artista,
 - O título da música,
 - A posição da música no *top 100* da *Billboard* em cada uma das 76 semanas depois que ela entrou na lista.
- Este último item é complexo, e o criador da *tibble* decidiu criar uma coluna por semana.
- Uma decisão ruim, pois existe informação embutida nos nomes das colunas. A coluna `wk68` corresponde à posição da música na semana 68 após ela entrar na lista, mas o número da semana só aparece no nome da coluna!
- Isto nunca deve acontecer. A informação deve sempre estar nas células.

- Vamos simplificar as coisas criando duas colunas:
 - semana, com o número da semana; perceba que esta informação vem dos nomes das colunas,
 - pos, com a posição da música naquela semana; esta informação vem das células.
- A *tibble*, que antes era larga, **vai ser mais estreita e mais longa.**
- A função `pivot_longer`, do pacote `tidyr`, vai fazer o trabalho — inclusive extraindo os números das semanas dos nomes das colunas:

```
bb_tidy <- bb %>%
  pivot_longer(
    wk1:wk76,
    names_to = 'semana',
    names_prefix = 'wk',
    names_transform = list(
      semana = as.integer
    ),
    values_to = 'pos'
  )

bb_tidy
```

```
## # A tibble: 24.092 x 5
##   artista musica      entrou  semana  pos
##   <chr>   <chr>      <date>    <int> <dbl>
## 1 2 Pac    Baby Don't Cry (Keep... 2000-02-26      1      87
## 2 2 Pac    Baby Don't Cry (Keep... 2000-02-26      2      82
## 3 2 Pac    Baby Don't Cry (Keep... 2000-02-26      3      72
## 4 2 Pac    Baby Don't Cry (Keep... 2000-02-26      4      77
## 5 2 Pac    Baby Don't Cry (Keep... 2000-02-26      5      87
## 6 2 Pac    Baby Don't Cry (Keep... 2000-02-26      6      94
## # ... with 24.086 more rows
```

- O R só mostra, por *default*, as 1000 primeiras linhas de uma *tibble*.
- Na verdade, o número de linhas da tabela original era

```
bb %>% nrow()
```

```
## [1] 317
```

- O número de linhas, depois de `pivot_longer`, ficou:

```
bb_tidy %>% nrow()
```

```
## [1] 24092
```

- Existem linhas onde `pos` tem o valor `NA`. São resultado da organização original dos dados, onde o `NA` indicava que a música não estava no *top 100* naquela semana.
- No novo formato, a ausência da linha com aquele número de semana já basta para indicar isto. Então, vamos eliminar as linhas onde `pos` é `NA`.
- A função `filter` **mantém** as linhas que **satisfazem** a condição dada; por isso, a condição é “`pos` não é `NA`”:

```
bb_tidy <- bb_tidy %>%
  filter(!is.na(pos))
```

```
bb_tidy
```

```
## # A tibble: 5.307 x 5
##   artista musica      entrou   semana   pos
##   <chr>    <chr>      <date>   <int> <dbl>
## 1 2 Pac    Baby Don't Cry (Keep... 2000-02-26     1     87
## 2 2 Pac    Baby Don't Cry (Keep... 2000-02-26     2     82
## 3 2 Pac    Baby Don't Cry (Keep... 2000-02-26     3     72
## 4 2 Pac    Baby Don't Cry (Keep... 2000-02-26     4     77
## 5 2 Pac    Baby Don't Cry (Keep... 2000-02-26     5     87
## 6 2 Pac    Baby Don't Cry (Keep... 2000-02-26     6     94
## # ... with 5.301 more rows
```

- O número de linhas ficou

```
bb_tidy %>% nrow()
```

```
## [1] 5307
```

3.3.1

Exercícios

- Todas as semanas deste conjunto de dados são do ano 2000?
- Qual é o tipo do **primeiro** argumento da função `filter()`?

3.4

Manipulando os dados

3.4.1

Criando novas colunas: mutate, transmute

- O *data frame*² `cars` tem dados (de 1920!) sobre as distâncias de frenagem (em pés) de um carro viajando a diversas velocidades (em milhas por hora):

```
cars
```

```
## # A tibble: 50 x 2
##   speed  dist
##   <dbl> <dbl>
## 1     4     2
## 2     4    10
## 3     7     4
## 4     7    22
## 5     8    16
## 6     9    10
## # ... with 44 more rows
```

- Vamos criar colunas novas com os valores convertidos para km/h e metros; além disso, uma coluna com a taxa de frenagem:

```
cars %>%
  mutate(
    velocidade = speed * 1.6,
    distancia = dist * .33,
    taxa = velocidade / distancia
  )
```

```
## # A tibble: 50 x 5
##   speed  dist velocidade distancia taxa
##   <dbl> <dbl>      <dbl>      <dbl> <dbl>
## 1     4     2        6.4        0.66  9.70
## 2     4    10        6.4        3.3   1.94
## 3     7     4       11.2        1.32  8.48
## 4     7    22       11.2        7.26  1.54
## 5     8    16       12.8        5.28  2.42
## 6     9    10       14.4        3.3   4.36
## # ... with 44 more rows
```

- Perceba que as colunas antigas continuam lá. Se quiser manter apenas as colunas novas, use `transmute`:

```
cars %>%
  transmute(
    velocidade = speed * 1.6,
```

²Considere *data frame* como sinônimo de *tibble*. Na verdade, *tibbles* formam um superconjunto de *data frames*: todo *data frame* é uma *tibble*, mas nem toda *tibble* é um *data frame*.

```

    distancia = dist * .33,
    taxa = velocidade / distancia
  )

```

```

## # A tibble: 50 x 3
##   velocidade distancia taxa
##   <dbl>      <dbl> <dbl>
## 1      6.4      0.66  9.70
## 2      6.4      3.3   1.94
## 3     11.2      1.32  8.48
## 4     11.2      7.26  1.54
## 5     12.8      5.28  2.42
## 6     14.4      3.3   4.36
## # ... with 44 more rows

```

- Ou use o argumento `.keep` de `mutate` para escolher com mais precisão. Veja a ajuda de `mutate`.

3.4.2

Selecionando colunas: `select`, `distinct`, `pull`

- Vamos voltar à nossa *tibble* dos *top 100* da *Billboard*.
- Para ver só a coluna de artistas:

```

bb_tidy %>%
  select(artista)

```

```

## # A tibble: 5.307 x 1
##   artista
##   <chr>
## 1 2 Pac
## 2 2 Pac
## 3 2 Pac
## 4 2 Pac
## 5 2 Pac
## 6 2 Pac
## # ... with 5.301 more rows

```

- Para eliminar as repetições:

```

bb_tidy %>%
  select(artista) %>%
  distinct()

```

```

## # A tibble: 228 x 1
##   artista
##   <chr>
## 1 2 Pac

```

```
## 2 2Ge+her
## 3 3 Doors Down
## 4 504 Boyz
## 5 98^0
## 6 A*Teens
## # ... with 222 more rows
```

- Para ver artistas e músicas:

```
bb_tidy %>%
  select(artista, musica) %>%
  distinct()
```

```
## # A tibble: 317 x 2
##   artista      musica
##   <chr>        <chr>
## 1 2 Pac        Baby Don't Cry (Keep...
## 2 2Ge+her      The Hardest Part Of ...
## 3 3 Doors Down Kryptonite
## 4 3 Doors Down Loser
## 5 504 Boyz     Wobble Wobble
## 6 98^0        Give Me Just One Nig...
## # ... with 311 more rows
```

- Para especificar colunas **a não mostrar**, use o sinal de menos “-”:

```
bb_tidy %>%
  select(-c(entrou, semana, pos))
```

```
## # A tibble: 5.307 x 2
##   artista musica
##   <chr>    <chr>
## 1 2 Pac    Baby Don't Cry (Keep...
## 2 2 Pac    Baby Don't Cry (Keep...
## 3 2 Pac    Baby Don't Cry (Keep...
## 4 2 Pac    Baby Don't Cry (Keep...
## 5 2 Pac    Baby Don't Cry (Keep...
## 6 2 Pac    Baby Don't Cry (Keep...
## # ... with 5.301 more rows
```

- Para **extrair uma coluna na forma de vetor** (unique é uma função do R base, aplicável a vetores):

```
bb_tidy %>%
  pull(artista) %>%
  unique()
```

```
##   [1] "2 Pac"                "2Ge+her"
##   [3] "3 Doors Down"         "504 Boyz"
```


##	[5]	"98~0"	"A*Teens"
##	[7]	"Aaliyah"	"Adams, Yolanda"
##	[9]	"Adkins, Trace"	"Aguilera, Christina"
##	[11]	"Alice DeeJay"	"Allan, Gary"
##	[13]	"Amber"	"Anastacia"
##	[15]	"Anthony, Marc"	"Avant"
##	[17]	"BBMak"	"Backstreet Boys, The"
##	[19]	"Badu, Erkyah"	"Baha Men"
##	[21]	"Barenaked Ladies"	"Beenie Man"
##	[23]	"Before Dark"	"Bega, Lou"
##	[25]	"Big Punisher"	"Black Rob"
##	[27]	"Black, Clint"	"Blaque"
##	[29]	"Blige, Mary J."	"Blink-182"
##	[31]	"Bloodhound Gang"	"Bon Jovi"
##	[33]	"Braxton, Toni"	"Brock, Chad"
##	[35]	"Brooks & Dunn"	"Brooks, Garth"
##	[37]	"Byrd, Tracy"	"Cagle, Chris"
##	[39]	"Cam'ron"	"Carey, Mariah"
##	[41]	"Carter, Aaron"	"Carter, Torrey"
##	[43]	"Changing Faces"	"Chesney, Kenny"
##	[45]	"Clark Family Experience"	"Clark, Terri"
##	[47]	"Common"	"Counting Crows"
##	[49]	"Creed"	"Cyrus, Billy Ray"
##	[51]	"D'Angelo"	"DMX"
##	[53]	"Da Brat"	"Davidson, Clay"
##	[55]	"De La Soul"	"Destiny's Child"
##	[57]	"Diffie, Joe"	"Dion, Celine"
##	[59]	"Dixie Chicks, The"	"Dr. Dre"
##	[61]	"Drama"	"Dream"
##	[63]	"Eastsidaz, The"	"Eiffel 65"
##	[65]	"Elliott, Missy \"Misdemeanor\""	"Eminem"
##	[67]	"En Vogue"	"Estefan, Gloria"
##	[69]	"Evans, Sara"	"Eve"
##	[71]	"Everclear"	"Fabian, Lara"
##	[73]	"Fatboy Slim"	"Filter"
##	[75]	"Foo Fighters"	"Fragma"
##	[77]	"Funkmaster Flex"	"Ghostface Killah"
##	[79]	"Gill, Vince"	"Gilman, Billy"
##	[81]	"Ginuwine"	"Goo Goo Dolls"
##	[83]	"Gray, Macy"	"Griggs, Andy"
##	[85]	"Guy"	"Hanson"
##	[87]	"Hart, Beth"	"Heatherly, Eric"
##	[89]	"Henley, Don"	"Herndon, Ty"
##	[91]	"Hill, Faith"	"Hoku"
##	[93]	"Hollister, Dave"	"Hot Boys"
##	[95]	"Houston, Whitney"	"IMx"
##	[97]	"Ice Cube"	"Ideal"
##	[99]	"Iglesias, Enrique"	"J-Shin"

## [101] "Ja Rule"	"Jackson, Alan"
## [103] "Jagged Edge"	"Janet"
## [105] "Jay-Z"	"Jean, Wyclef"
## [107] "Joe"	"John, Elton"
## [109] "Jones, Donell"	"Jordan, Montell"
## [111] "Juvenile"	"Kandi"
## [113] "Keith, Toby"	"Kelis"
## [115] "Kenny G"	"Kid Rock"
## [117] "Kravitz, Lenny"	"Kumbia Kings"
## [119] "LFO"	"LL Cool J"
## [121] "Larrieux, Amel"	"Lawrence, Tracy"
## [123] "Levert, Gerald"	"Lil Bow Wow"
## [125] "Lil Wayne"	"Lil' Kim"
## [127] "Lil' Mo"	"Lil' Zane"
## [129] "Limp Bizkit"	"Lonestar"
## [131] "Lopez, Jennifer"	"Loveless, Patty"
## [133] "Lox"	"Lucy Pearl"
## [135] "Ludacris"	"M2M"
## [137] "Madison Avenue"	"Madonna"
## [139] "Martin, Ricky"	"Mary Mary"
## [141] "Master P"	"McBride, Martina"
## [143] "McEntire, Reba"	"McGraw, Tim"
## [145] "McKnight, Brian"	"Messina, Jo Dee"
## [147] "Metallica"	"Montgomery Gentry"
## [149] "Montgomery, John Michael"	"Moore, Chante"
## [151] "Moore, Mandy"	"Mumba, Samantha"
## [153] "MusiQ"	"Mya"
## [155] "Mystikal"	"N'Sync"
## [157] "Nas"	"Nelly"
## [159] "Next"	"Nine Days"
## [161] "No Doubt"	"Nu Flavor"
## [163] "Offspring, The"	"Paisley, Brad"
## [165] "Papa Roach"	"Pearl Jam"
## [167] "Pink"	"Price, Kelly"
## [169] "Profyle"	"Puff Daddy"
## [171] "Q-Tip"	"R.E.M."
## [173] "Rascal Flatts"	"Raye, Collin"
## [175] "Red Hot Chili Peppers"	"Rimes, LeAnn"
## [177] "Rogers, Kenny"	"Ruff Endz"
## [179] "Sammie"	"Santana"
## [181] "Savage Garden"	"SheDaisy"
## [183] "Sheist, Shade"	"Shyne"
## [185] "Simpson, Jessica"	"Sisqo"
## [187] "Sister Hazel"	"Smash Mouth"
## [189] "Smith, Will"	"Son By Four"
## [191] "Sonique"	"SoulDecision"
## [193] "Spears, Britney"	"Spencer, Tracie"
## [195] "Splendor"	"Sting"

## [197]	"Stone Temple Pilots"	"Stone, Angie"
## [199]	"Strait, George"	"Sugar Ray"
## [201]	"TLC"	"Tamar"
## [203]	"Tamia"	"Third Eye Blind"
## [205]	"Thomas, Carl"	"Tippin, Aaron"
## [207]	"Train"	"Trick Daddy"
## [209]	"Trina"	"Tritt, Travis"
## [211]	"Tuesday"	"Urban, Keith"
## [213]	"Usher"	"Vassar, Phil"
## [215]	"Vertical Horizon"	"Vitamin C"
## [217]	"Walker, Clay"	"Wallflowers, The"
## [219]	"Westlife"	"Williams, Robbie"
## [221]	"Wills, Mark"	"Worley, Darryl"
## [223]	"Wright, Chely"	"Yankee Grey"
## [225]	"Yearwood, Trisha"	"Ying Yang Twins"
## [227]	"Zombie Nation"	"matchbox twenty"

3.4.3

Filtrando linhas: `filter`, `slice`

- Apenas as músicas da Britney Spears:

```
bb_tidy %>%
  filter(artista == 'Spears, Britney')
```

```
## # A tibble: 51 x 5
##   artista      musica      entrou      semana  pos
##   <chr>        <chr>        <date>      <int> <dbl>
## 1 Spears, Britney From The Bottom Of M... 2000-01-29      1    76
## 2 Spears, Britney From The Bottom Of M... 2000-01-29      2    59
## 3 Spears, Britney From The Bottom Of M... 2000-01-29      3    52
## 4 Spears, Britney From The Bottom Of M... 2000-01-29      4    52
## 5 Spears, Britney From The Bottom Of M... 2000-01-29      5    14
## 6 Spears, Britney From The Bottom Of M... 2000-01-29      6    14
## # ... with 45 more rows
```

- Apenas músicas que chegaram à posição 1, sem mostrar a coluna pos:

```
bb_tidy %>%
  filter(pos == 1) %>%
  select(-pos)
```

```
## # A tibble: 55 x 4
##   artista      musica      entrou      semana
##   <chr>        <chr>        <date>      <int>
## 1 Aaliyah      Try Again    2000-03-18      14
## 2 Aguilera, Christina Come On Over Baby (A... 2000-08-05      11
```

```
## 3 Aguilera, Christina Come On Over Baby (A... 2000-08-05      12
## 4 Aguilera, Christina Come On Over Baby (A... 2000-08-05      13
## 5 Aguilera, Christina Come On Over Baby (A... 2000-08-05      14
## 6 Aguilera, Christina What A Girl Wants      1999-11-27      8
## # ... with 49 more rows
```

- Apenas músicas que chegaram à posição 1 em menos de 10 semanas, mostrando apenas artista e música:

```
bb_tidy %>%
  filter(pos == 1, semana < 10) %>%
  distinct(artista, musica)
```

```
## # A tibble: 5 x 2
##   artista      musica
##   <chr>        <chr>
## 1 Aguilera, Christina What A Girl Wants
## 2 Destiny's Child Independent Women Pa...
## 3 Madonna      Music
## 4 Santana      Maria, Maria
## 5 Sisqo        Incomplete
```

- As funções da família `slice` filtram linhas de diversas maneiras.
- De acordo com seus índices (números de linha):

```
bb_tidy %>%
  slice(c(1, 1000, 5000))
```

```
## # A tibble: 3 x 5
##   artista      musica      entrou      semana  pos
##   <chr>        <chr>      <date>      <int> <dbl>
## 1 2 Pac        Baby Don't Cry (Keep~ 2000-02-26      1    87
## 2 Clark Family Experience Meanwhile Back At Th~ 2000-11-18      3    81
## 3 Vassar, Phil Carlene      2000-03-04      3    64
```

```
bb_tidy %>%
  slice_head(n = 4)
```

```
## # A tibble: 4 x 5
##   artista musica      entrou      semana  pos
##   <chr>   <chr>      <date>      <int> <dbl>
## 1 2 Pac   Baby Don't Cry (Keep... 2000-02-26      1    87
## 2 2 Pac   Baby Don't Cry (Keep... 2000-02-26      2    82
## 3 2 Pac   Baby Don't Cry (Keep... 2000-02-26      3    72
## 4 2 Pac   Baby Don't Cry (Keep... 2000-02-26      4    77
```

```
bb_tidy %>%
  slice_tail(n = 4)
```

```
## # A tibble: 4 x 5
##   artista      musica entrou   semana   pos
##   <chr>        <chr>   <date>     <int> <dbl>
## 1 matchbox twenty Bent    2000-04-29     36    37
## 2 matchbox twenty Bent    2000-04-29     37    38
## 3 matchbox twenty Bent    2000-04-29     38    38
## 4 matchbox twenty Bent    2000-04-29     39    48
```

- De acordo com a ordenação de uma coluna ou de uma função das colunas:

```
bb_tidy %>%
  slice_min(pos)
```

```
## # A tibble: 55 x 5
##   artista      musica      entrou   semana   pos
##   <chr>        <chr>    <date>     <int> <dbl>
## 1 Aaliyah      Try Again 2000-03-18     14     1
## 2 Aguilera, Christina Come On Over Baby (A... 2000-08-05     11     1
## 3 Aguilera, Christina Come On Over Baby (A... 2000-08-05     12     1
## 4 Aguilera, Christina Come On Over Baby (A... 2000-08-05     13     1
## 5 Aguilera, Christina Come On Over Baby (A... 2000-08-05     14     1
## 6 Aguilera, Christina What A Girl Wants      1999-11-27      8     1
## # ... with 49 more rows
```

```
bb_tidy %>%
  slice_max(semana)
```

```
## # A tibble: 1 x 5
##   artista musica      entrou   semana   pos
##   <chr>   <chr>   <date>     <int> <dbl>
## 1 Creed  Higher 1999-09-11     65    49
```

- Aleatoriamente, criando uma amostra:

```
bb_tidy %>%
  slice_sample(n = 5)
```

```
## # A tibble: 5 x 5
##   artista      musica      entrou   semana   pos
##   <chr>        <chr>    <date>     <int> <dbl>
## 1 Cagle, Chris  My Love Goes On And ... 2000-10-21     10    83
## 2 Hollister, Dave Can't Stay      2000-03-25      2    84
## 3 Price, Kelly  Love Sets You Free    2000-05-13      4   100
## 4 Lox           Ryde or Die, Chick    2000-03-18      2    73
## 5 Backstreet Boys, The Shape Of My Heart    2000-10-14      2    25
```

- Veja a ajuda de `slice` para saber mais sobre estas funções. Por exemplo:
 - `slice_min` e `slice_max` podem considerar ou não empates.
 - Você pode especificar uma proporção de linhas (usando `prop`) em vez da quantidade de linhas (`n`).
 - Você pode fazer amostragem com reposição, ou com probabilidades diferentes para cada linha.

3.4.4

Ordenando linhas: `arrange`

- Por título, sem repetições:

```
bb_tidy %>%
  select(musica) %>%
  distinct() %>%
  arrange(musica)
```

```
## # A tibble: 316 x 1
##   musica
##   <chr>
## 1 (Hot S**t) Country G...
## 2 3 Little Words
## 3 911
## 4 A Country Boy Can Su...
## 5 A Little Gasoline
## 6 A Puro Dolor (Purest...
## # ... with 310 more rows
```

- Por título, sem repetições, em ordem inversa:

```
bb_tidy %>%
  select(musica) %>%
  distinct() %>%
  arrange(desc(musica))
```

```
## # A tibble: 316 x 1
##   musica
##   <chr>
## 1 Your Everything
## 2 You're A God
## 3 You'll Always Be Lov...
## 4 You Won't Be Lonely ...
## 5 You Should've Told M...
## 6 You Sang To Me
## # ... with 310 more rows
```

3.4.5

Contando linhas: `count`

- Quantas semanas cada artista ficou nos *top 100*? Duas músicas na mesma semana contam como duas semanas.

```
bb_tidy %>%  
  count(artista, sort = TRUE)
```

```
## # A tibble: 228 x 2  
##   artista      n  
##   <chr>      <int>  
## 1 Creed      104  
## 2 Lonestar    95  
## 3 Destiny's Child 92  
## 4 N'Sync      74  
## 5 Sisqo       74  
## 6 3 Doors Down 73  
## # ... with 222 more rows
```

- Quantas semanas cada música ficou nos *top 100*?

```
bb_tidy %>%  
  count(musica, sort = TRUE)
```

```
## # A tibble: 316 x 2  
##   musica      n  
##   <chr>      <int>  
## 1 Higher     57  
## 2 Amazed     55  
## 3 Breathe    53  
## 4 Kryptonite 53  
## 5 With Arms Wide Open 47  
## 6 I Wanna Know 44  
## # ... with 310 more rows
```

- Se houve músicas com o mesmo nome, mas de artistas diferentes, o código acima está errado. O certo é

```
bb_tidy %>%  
  count(musica, artista, sort = TRUE)
```

```
## # A tibble: 317 x 3  
##   musica      artista      n  
##   <chr>      <chr>      <int>  
## 1 Higher     Creed        57  
## 2 Amazed     Lonestar     55  
## 3 Breathe    Hill, Faith  53
```

```
## 4 Kryptonite          3 Doors Down    53
## 5 With Arms Wide Open Creed          47
## 6 I Wanna Know       Joe              44
## # ... with 311 more rows
```

De fato, há uma diferença de uma linha.

3.4.5.1

Exercício

- Ache o título da música que tem dois artistas diferentes.

Sugestão: conte por música e artista primeiro, depois só por música.

3.4.6

Agrupando linhas: `group_by` e `summarize`

- Qual foi a melhor posição que cada artista alcançou?

```
bb_tidy %>%
  group_by(artista) %>%
  summarize(melhor = min(pos)) %>%
  arrange(melhor)
```

```
## # A tibble: 228 x 2
##   artista          melhor
##   <chr>            <dbl>
## 1 Aaliyah          1
## 2 Aguilera, Christina 1
## 3 Carey, Mariah     1
## 4 Creed            1
## 5 Destiny's Child   1
## 6 Iglesias, Enrique 1
## # ... with 222 more rows
```

- Qual foi a melhor posição que cada música alcançou?

```
bb_tidy %>%
  group_by(artista, musica) %>%
  summarize(melhor = min(pos)) %>%
  arrange(melhor)
```

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

```
## # A tibble: 317 x 3
##   artista          musica          melhor
##   <chr>            <chr>            <dbl>
```



```
## 1 Aaliyah          Try Again          1
## 2 Aguilera, Christina Come On Over Baby (A... 1
## 3 Aguilera, Christina What A Girl Wants      1
## 4 Carey, Mariah    Thank God I Found Yo... 1
## 5 Creed            With Arms Wide Open    1
## 6 Destiny's Child  Independent Women Pa... 1
## # ... with 311 more rows
```

- Quando usamos `summarize`, só o agrupamento **mais interno** é desfeito. Isto significa que **o resultado acima ainda está agrupado por artista**.
- Quantas semanas cada artista ficou na posição 1?

A função `n()` é uma maneira conveniente de **obter o número de linhas de um grupo** (ou, se não houver agrupamento, de toda a *tibble*); mas `n()` só pode ser chamada **em certos contextos**, como `summarise()` ou `mutate()`.

```
bb_tidy %>%
  filter(pos == 1) %>%
  group_by(artista) %>%
  summarize(semanas = n()) %>%
  arrange(desc(semanas))
```

```
## # A tibble: 15 x 2
##   artista          semanas
##   <chr>            <int>
## 1 Destiny's Child      14
## 2 Santana             10
## 3 Aguilera, Christina   6
## 4 Madonna             4
## 5 Savage Garden        4
## 6 Iglesias, Enrique     3
## # ... with 9 more rows
```

- Perceba que `count`, que vimos mais acima, faz agrupamentos do mesmo modo:

```
bb_tidy %>%
  filter(pos == 1) %>%
  count(artista, sort = TRUE)
```

```
## # A tibble: 15 x 2
##   artista          n
##   <chr>            <int>
## 1 Destiny's Child      14
## 2 Santana             10
## 3 Aguilera, Christina   6
## 4 Madonna             4
## 5 Savage Garden        4
## 6 Iglesias, Enrique     3
```

```
## # ... with 9 more rows
```

- Uma pergunta diferente: quais são os artistas cujas músicas apareceram no *top 100* mais tempo depois do lançamento da música?

```
bb_tidy %>%
  group_by(artista) %>%
  summarize(semanas = max(semana)) %>%
  arrange(desc(semanas))
```

```
## # A tibble: 228 x 2
##   artista      semanas
##   <chr>         <int>
## 1 Creed          65
## 2 Lonestar       64
## 3 3 Doors Down   53
## 4 Hill, Faith    53
## 5 Joe            44
## 6 Vertical Horizon 41
## # ... with 222 more rows
```

- Qual a posição média de cada música? Lembre-se de que eliminamos as linhas com NA; logo, **a média vai ser sobre a quantidade de semanas em que a música esteve na lista.**

```
media1 <- bb_tidy %>%
  group_by(artista, musica) %>%
  summarize(media = mean(pos), .groups = 'drop') %>%
  arrange(media)
```

```
media1
```

```
## # A tibble: 317 x 3
##   artista      musica      media
##   <chr>         <chr>    <dbl>
## 1 "Santana"      Maria, Maria    10.5
## 2 "Madonna"      Music          13.5
## 3 "N'Sync"       Bye Bye Bye     14.3
## 4 "Elliott, Missy \"Misdemeanor\"" Hot Boyz        14.3
## 5 "Destiny's Child" Independent Women Pa... 14.8
## 6 "Iglesias, Enrique" Be With You     15.8
## # ... with 311 more rows
```

- E se quisermos **a média sobre o número de semanas desde a entrada da música até a última semana** em que a música apareceu na lista?

```
media2 <- bb_tidy %>%
```

```
group_by(artista, musica) %>%
  summarize(media = sum(pos)/max(semana), .groups = 'drop') %>%
  arrange(media)
```

```
media2
```

```
## # A tibble: 317 x 3
##   artista          musica          media
##   <chr>          <chr>          <dbl>
## 1 "Santana"      Maria, Maria      10.5
## 2 "Madonna"     Music             13.5
## 3 "N'Sync"      Bye Bye Bye       14.3
## 4 "Elliott, Missy \"Misdemeanor\"" Hot Boyz          14.3
## 5 "Destiny's Child" Independent Women Pa... 14.8
## 6 "Iglesias, Enrique" Be With You       15.8
## # ... with 311 more rows
```

As primeiras linhas são iguais, mas os resultados são diferentes:

```
identical(media1, media2)
```

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

3.5

Exercícios

1. Vamos trabalhar com um conjunto de dados sobre super-heróis.

Carregue o tidyverse com o comando

```
library(tidyverse)
```

Execute o seguinte comando para ler os dados para uma *tibble*:

```
arquivo <- paste0(
  'https://github.com/fnaufel/',
  'probestr/raw/master/data/',
  'heroes_information.csv'
)

herois_info <- read_csv(
  arquivo,
  na = c('', '-', 'NA')
) %>%
  # Eliminar a primeira coluna (números de série)
  select(-1) %>%
  # Renomear colunas restantes
```

```

rename(
  nome = name,
  sexo = Gender,
  olhos = 'Eye color',
  raça = Race,
  cabelos = 'Hair color',
  altura = Height,
  editora = Publisher,
  pele = 'Skin color',
  lado = Alignment,
  peso = Weight
)

```

2. Quantas linhas tem a *tibble*?
3. Existem heróis que aparecem em mais de uma linha?
4. Quantas editoras diferentes existem na *tibble*? Liste-as em ordem decrescente de quantidade de heróis.
5. Vamos colocar todas as editores menores em uma classe só.

Na coluna *editora*, substitua

- ‘Marvel Comics’ por ‘Marvel’,
- ‘DC Comics’ por ‘DC’, e
- todas as outras editoras pelo termo ‘Outras’.

Dica: use a função `case_when()`, do tidyverse.

6. Confira, novamente, a quantidade de valores diferentes na coluna *editora*.
7. Existem heróis sem informação de editora. Quantos? Quais são?
8. Altere novamente a coluna *editora*, colocando o valor ‘Outras’ para os heróis sem informação de editora. Use a função `if_else()` (com *underscore*, não a função `ifelse`).
9. Confira, mais uma vez, a quantidade de valores diferentes na coluna *editora*.
10. Existem heróis sem informação de sexo? Quantos? Para estes heróis, coloque o valor ‘Desconhecido’ na coluna *sexo*.
11. Qual a altura mínima? Qual a altura máxima? Substitua as alturas negativas por `NA`.
12. Qual o peso mínimo? Qual o peso máximo? Substitua os pesos negativos por `NA`.
13. Qual é o peso médio de todos os heróis? Ignore os valores `NA`.
14. Qual é a altura média de todos os heróis? Ignore os valores `NA`.
15. Qual é a altura média dos heróis, por editora? Ignore os valores `NA`.
16. Quais são os 3 heróis mais altos de cada sexo?
17. Quais são as 3 cores de olhos mais comuns para cada sexo?
18. Liste, por editora, as quantidades de heróis do bem, do mal, e neutros.

19. Quantas raças diferentes existem?

20. Qual a quantidade de raças diferentes de cada editora?

21. **DESAFIO:** Liste as raças que só pertencem a uma única editora.

Existem várias maneiras de fazer isto. Experimente várias, até achar uma que seja mais elegante.

Visualização com ggplot2



Busque mais informações sobre os pacotes `tidyverse` e `ggplot2` nas referências recomendadas.

4.1

Vídeo 1

<https://youtu.be/OBpNjqIIyhI>

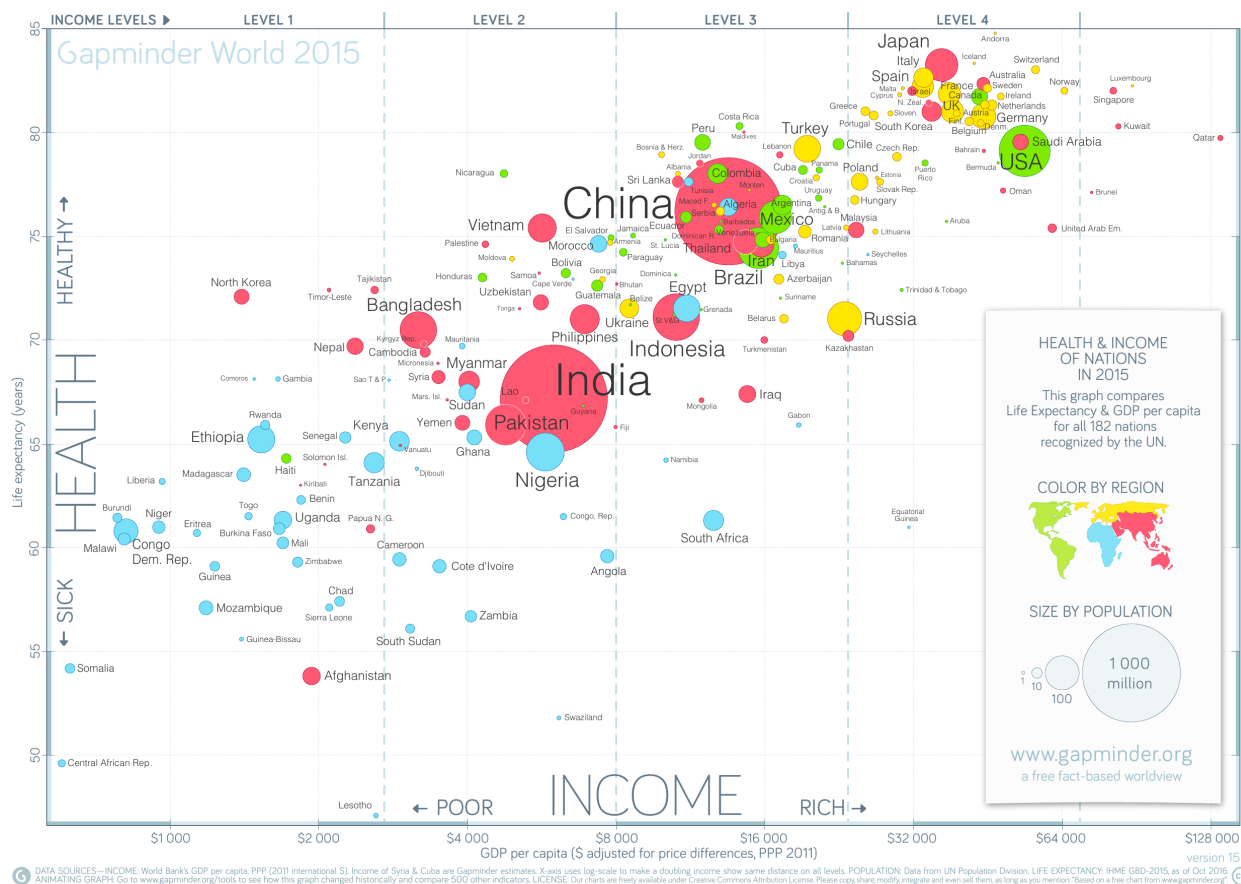
4.2

Componentes de um gráfico ggplot2

4.2.1

Geometrias e mapeamentos estéticos (*mappings*)

- Observe o gráfico abaixo, obtido de <https://www.gapminder.org/downloads/updated-gapminder-world-poster-2015/>.



- O gráfico mostra como, em cada país, a saúde (mais precisamente, a expectativa de vida) se relaciona com a riqueza (mais precisamente, o PIB *per capita*).
- Além da expectativa de vida e o do PIB *per capita*, o gráfico traz mais informações sobre cada país.
- Cada país é representado por um ponto (a **geometria**).
- Informações sobre cada país são representadas por características do ponto correspondente (as **estéticas**):

Variável	Geometria	Estética
PIB <i>per capita</i>	ponto	posição x
Expectativa de vida	ponto	posição y
População	ponto	tamanho
Continente	ponto	cor

- Você pode usar outras estéticas para representar informações:
 - Cor de preenchimento.
 - Cor do traço.
 - Tipo do traço (sólido, pontilhado, tracejado etc.).
 - Forma (círculo, quadrado, triângulo etc.).

- Opacidade.
- etc.
- Você pode usar outras geometrias:
 - Linhas.
 - Barras ou colunas.
 - Caixas.
 - etc.

4.2.2

Escalas (*scales*)

- As escalas controlam os detalhes da aparência da geometria e do mapeamento (eixos, cores etc.).
- Os eixos do gráfico acima são escalas **contínuas**, com valores reais.
- Observe o eixo horizontal. Os valores não aumentam linearmente, mas sim exponencialmente: cada passo à direita equivale a *dobrar* o valor do PIB. O eixo horizontal segue uma **escala logarítmica**.
- Os tamanhos dos pontos formam uma escala **discreta**, com 4 valores possíveis (veja a legenda no canto inferior direito do gráfico).
- As cores também formam uma escala discreta.

4.2.3

Rótulos (*labels*)

- O gráfico também representa informação na forma de texto.
- Além de rótulos (por exemplo, o texto que identifica cada eixo), **o texto também pode, ele mesmo, ser uma geometria, com suas próprias estéticas**: observe como o nome de cada país é escrito em um tamanho proporcional à sua população.

4.2.4

Outros componentes

- Coordenadas:
 - Este gráfico usa **coordenadas cartesianas**, com eixos x e y .
 - Existem gráficos que usam um sistema de **coordenadas polares**.
- Temas:
 - Incluem todos os elementos “decorativos”: cor de fundo, linhas de grade, etc. Ajudam a facilitar a leitura e a interpretação.

- No gráfico acima, um detalhe interessante do tema é a divisão de cada eixo em segmentos claros e segmentos escuros.
- Legendas (*guides*).
- Facetas:
 - Às vezes, um gráfico é composto por múltiplos subgráficos.
 - Cada subgráfico é uma **faceta**.
 - Facetas evitam que informações demais sejam apresentadas no mesmo lugar.

4.3

Conjunto de dados

- Nossos exemplos de gráficos vão usar dados sobre o sono de diversos mamíferos.
- O conjunto de dados se chama `msleep` e está incluído no pacote `ggplot2`.
- Para ver a documentação, digite

```
library(ggplot2)
?msleep
```

- Vamos atribuir o conjunto de dados à variável `df`:

```
df <- msleep
df
```

```
## # A tibble: 83 x 11
##   name          genus vore  order conservation sleep_total sleep_rem
##   <chr>         <chr> <chr> <chr> <chr>           <dbl>     <dbl>
## 1 Cheetah      Acin~ carni Carn~ lc             12.1      NA
## 2 Owl monkey   Aotus omni Prim~ <NA>         17        1.8
## 3 Mountain beaver Aplo~ herbi Rode~ nt          14.4      2.4
## 4 Greater short-t~ Blar~ omni Sori~ lc          14.9      2.3
## 5 Cow          Bos   herbi Arti~ domesticated    4        0.7
## 6 Three-toed sloth Brad~ herbi Pilo~ <NA>         14.4      2.2
## # ... with 77 more rows, and 4 more variables: sleep_cycle <dbl>,
## #   awake <dbl>, brainwt <dbl>, bodywt <dbl>
```

- Vamos examinar a estrutura — usando R base:

```
str(df)
```

```
## tibble [83 x 11] (S3: tbl_df/tbl/data.frame)
##  $ name          : chr [1:83] "Cheetah" "Owl monkey" "Mountain beaver" ...
##  $ genus         : chr [1:83] "Acinonyx" "Aotus" "Aplodontia" ...
##  $ vore          : chr [1:83] "carni" "omni" "herbi" ...
```

```
## $ order      : chr [1:83] "Carnivora" "Primates" "Rodentia" ...
## $ conservation: chr [1:83] "lc" NA "nt" ...
## $ sleep_total : num [1:83] 12,1 17 14,4 14,9 4 14,4 8,7 7 ...
## $ sleep_rem   : num [1:83] NA 1,8 2,4 2,3 0,7 2,2 1,4 NA ...
## $ sleep_cycle : num [1:83] NA NA NA 0,133 ...
## $ awake       : num [1:83] 11,9 7 9,6 9,1 20 9,6 15,3 17 ...
## $ brainwt     : num [1:83] NA 0,0155 NA 0,00029 0,423 NA NA NA ...
## $ bodywt      : num [1:83] 50 0,48 1,35 0,019 ...
```

- Podemos usar `glimpse`, uma função do tidyverse:

```
glimpse(df)
```

```
## Rows: 83
## Columns: 11
## $ name      <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Gre~
## $ genus     <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "B~
## $ vore      <chr> "carni", "omni", "herbi", "omni", "herbi", "herb~
## $ order     <chr> "Carnivora", "Primates", "Rodentia", "Soricomorp~
## $ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", ~
## $ sleep_total <dbl> 12,1, 17,0, 14,4, 14,9, 4,0, 14,4, 8,7, 7,0, 10,~
## $ sleep_rem  <dbl> NA, 1,8, 2,4, 2,3, 0,7, 2,2, 1,4, NA, 2,9, NA, 0~
## $ sleep_cycle <dbl> NA, NA, NA, 0,1333333, 0,6666667, 0,7666667, 0,3~
## $ awake     <dbl> 11,9, 7,0, 9,6, 9,1, 20,0, 9,6, 15,3, 17,0, 13,9~
## $ brainwt   <dbl> NA, 0,01550, NA, 0,00029, 0,42300, NA, NA, NA, 0~
## $ bodywt    <dbl> 50,000, 0,480, 1,350, 0,019, 600,000, 3,850, 20,~
```

- Para examinar só as primeiras linhas do *data frame*:

```
head(df)
```

```
## # A tibble: 6 x 11
##   name      genus vore order conservation sleep_total sleep_rem
##   <chr>      <chr> <chr> <chr> <chr>          <dbl>      <dbl>
## 1 Cheetah   Acin~ carn~ Carn~ lc             12.1        NA
## 2 Owl monkey Aotus omni  Prim~ <NA>          17          1.8
## 3 Mountain beaver Aplo~ herbi Rode~ nt             14.4        2.4
## 4 Greater short-t~ Blar~ omni  Sori~ lc             14.9        2.3
## 5 Cow       Bos   herbi Arti~ domesticated    4          0.7
## 6 Three-toed sloth Brad~ herbi Pilo~ <NA>          14.4        2.2
## # ... with 4 more variables: sleep_cycle <dbl>, awake <dbl>,
## #   brainwt <dbl>, bodywt <dbl>
```

- Para examinar o *data frame* interativamente:

```
view(df)
```

- Podemos produzir um sumário dos dados usando o pacote *summarytools* (que já foi carregado neste documento):

```
df %>% dfSummary() %>% print()
```

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
name	1. African elephant	1 (1,2%)	0
[character]	2. African giant pouched rat	1 (1,2%)	(0,0%)
	3. African striped mouse	1 (1,2%)	
	4. Arctic fox	1 (1,2%)	
	5. Arctic ground squirrel	1 (1,2%)	
	6. Asian elephant	1 (1,2%)	
	7. Baboon	1 (1,2%)	
	8. Big brown bat	1 (1,2%)	
	9. Bottle-nosed dolphin	1 (1,2%)	
	10. Brazilian tapir	1 (1,2%)	
	[73 outros]	73 (88,0%)	
genus	1. Panthera	3 (3,6%)	0
[character]	2. Spermophilus	3 (3,6%)	(0,0%)
	3. Equus	2 (2,4%)	
	4. Vulpes	2 (2,4%)	
	5. Acinonyx	1 (1,2%)	
	6. Aotus	1 (1,2%)	
	7. Aplodontia	1 (1,2%)	
	8. Blarina	1 (1,2%)	
	9. Bos	1 (1,2%)	
	10. Bradypus	1 (1,2%)	
	[67 outros]	67 (80,7%)	
vore	1. carni	19 (25,0%)	7
[character]	2. herbi	32 (42,1%)	(8,4%)
	3. insecti	5 (6,6%)	
	4. omni	20 (26,3%)	
order	1. Rodentia	22 (26,5%)	0
[character]	2. Carnivora	12 (14,5%)	(0,0%)
	3. Primates	12 (14,5%)	
	4. Artiodactyla	6 (7,2%)	
	5. Soricomorpha	5 (6,0%)	
	6. Cetacea	3 (3,6%)	
	7. Hyracoidea	3 (3,6%)	
	8. Perissodactyla	3 (3,6%)	
	9. Chiroptera	2 (2,4%)	
	10. Cingulata	2 (2,4%)	
	[9 outros]	13 (15,7%)	
conservation	1. cd	2 (3,7%)	29
[character]	2. domesticated	10 (18,5%)	(34,9%)
	3. en	4 (7,4%)	
	4. lc	27 (50,0%)	
	5. nt	4 (7,4%)	
	6. vu	7 (13,0%)	

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
sleep_total [numeric]	Média (dp) : 10,4 (4,5) mín < mediana < máx: 1,9 < 10,1 < 19,9 IQE (CV) : 5,9 (0,4)	65 valores distintos	0 (0,0%)
sleep_rem [numeric]	Média (dp) : 1,9 (1,3) mín < mediana < máx: 0,1 < 1,5 < 6,6 IQE (CV) : 1,5 (0,7)	32 valores distintos	22 (26,5%)
sleep_cycle [numeric]	Média (dp) : 0,4 (0,4) mín < mediana < máx: 0,1 < 0,3 < 1,5 IQE (CV) : 0,4 (0,8)	22 valores distintos	51 (61,4%)
awake [numeric]	Média (dp) : 13,6 (4,5) mín < mediana < máx: 4,1 < 13,9 < 22,1 IQE (CV) : 5,9 (0,3)	65 valores distintos	0 (0,0%)
brainwt [numeric]	Média (dp) : 0,3 (1) mín < mediana < máx: 0 < 0 < 5,7 IQE (CV) : 0,1 (3,5)	53 valores distintos	27 (32,5%)
bodywt [numeric]	Média (dp) : 166,1 (786,8) mín < mediana < máx: 0 < 1,7 < 6654 IQE (CV) : 41,6 (4,7)	82 valores distintos	0 (0,0%)

- Vemos que há muitos NA em diversas variáveis. Para nossos exemplos simples de visualização, vamos usar as colunas

```
- name
- genus
- order
- sleep_total
- awake
- bodywt
- brainwt
```

- Mas... a coluna que mostra a dieta (vore) tem só 7 NA. Quais são?

```
df %>%
  filter(is.na(vore)) %>%
  select(name)
```

```
## # A tibble: 7 x 1
##   name
##   <chr>
## 1 Vesper mouse
```

```
## 2 Desert hedgehog
## 3 Deer mouse
## 4 Phalanger
## 5 Rock hyrax
## 6 Mole rat
## # ... with 1 more row
```

- OK. Vamos manter a coluna `vore` também, apesar dos NA. Quando formos usar esta variável, tomaremos cuidado.
- Também... a coluna `bodywt` tem 0 como valor mínimo. Como assim?

```
df %>%
  filter(bodywt < 1) %>%
  select(name, bodywt) %>%
  arrange(bodywt)
```

```
## # A tibble: 35 x 2
##   name                bodywt
##   <chr>              <dbl>
## 1 Lesser short-tailed shrew 0.005
## 2 Little brown bat        0.01
## 3 Greater short-tailed shrew 0.019
## 4 Deer mouse              0.021
## 5 House mouse             0.022
## 6 Big brown bat           0.023
## # ... with 29 more rows
```

- Ah, sem problema. A função `dfSummary` arredondou estes pesos para 0. Os valores de verdade ainda estão na *tibble*.
- Vamos criar uma *tibble* nova, só com as colunas que nos interessam:

```
sono <- df %>%
  select(
    name, order, genus, vore, bodywt,
    brainwt, awake, sleep_total
  )
```

- Vamos ver o sumário:

```
sono %>% dfSummary() %>% print()
```

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
name [character]	1. African elephant 2. African giant pouched rat 3. African striped mouse 4. Arctic fox 5. Arctic ground squirrel 6. Asian elephant 7. Baboon 8. Big brown bat 9. Bottle-nosed dolphin 10. Brazilian tapir [73 outros]	1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 73 (88,0%)	0 (0,0%)
order [character]	1. Rodentia 2. Carnivora 3. Primates 4. Artiodactyla 5. Soricomorpha 6. Cetacea 7. Hyracoidea 8. Perissodactyla 9. Chiroptera 10. Cingulata [9 outros]	22 (26,5%) 12 (14,5%) 12 (14,5%) 6 (7,2%) 5 (6,0%) 3 (3,6%) 3 (3,6%) 3 (3,6%) 2 (2,4%) 2 (2,4%) 13 (15,7%)	0 (0,0%)
genus [character]	1. Panthera 2. Sperophilus 3. Equus 4. Vulpes 5. Acinonyx 6. Aotus 7. Aplodontia 8. Blarina 9. Bos 10. Bradypus [67 outros]	3 (3,6%) 3 (3,6%) 2 (2,4%) 2 (2,4%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 67 (80,7%)	0 (0,0%)
vore [character]	1. carni 2. herbi 3. insecti 4. omni	19 (25,0%) 32 (42,1%) 5 (6,6%) 20 (26,3%)	7 (8,4%)
bodywt [numeric]	Média (dp) : 166,1 (786,8) mín < mediana < máx: 0 < 1,7 < 6654 IQE (CV) : 41,6 (4,7)	82 valores distintos	0 (0,0%)
brainwt [numeric]	Média (dp) : 0,3 (1) mín < mediana < máx: 0 < 0 < 5,7 IQE (CV) : 0,1 (3,5)	53 valores distintos	27 (32,5%)

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
awake [numeric]	Média (dp) : 13,6 (4,5) mín < mediana < máx: 4,1 < 13,9 < 22,1 IQE (CV) : 5,9 (0,3)	65 valores distintos	0 (0,0%)
sleep_total [numeric]	Média (dp) : 10,4 (4,5) mín < mediana < máx: 1,9 < 10,1 < 19,9 IQE (CV) : 5,9 (0,4)	65 valores distintos	0 (0,0%)

4.4

Gráficos de dispersão (*scatter plots*)

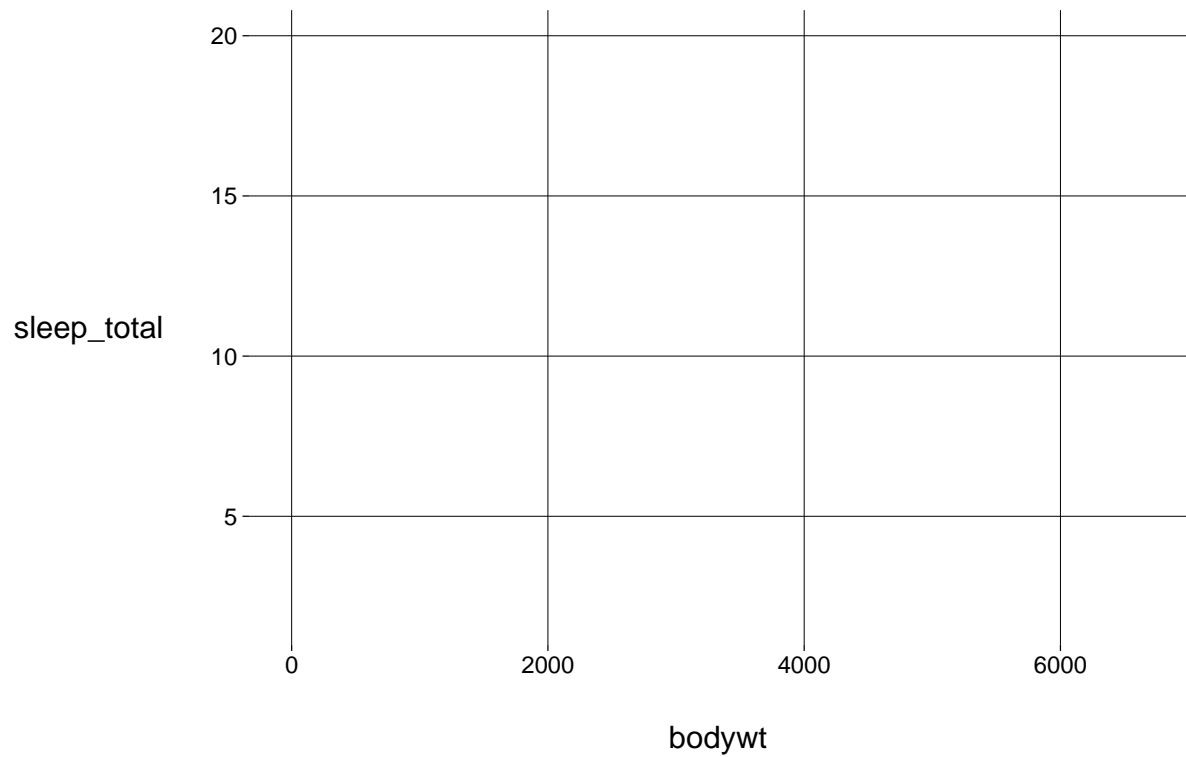
- Servem para visualizar a *relação* entre **duas variáveis quantitativas**.
- **Essa relação não é necessariamente de causa e efeito**.
- Isto é, a variável do eixo horizontal não determina, necessariamente, os valores da variável do eixo vertical.
- Pense em **associação**, **correlação**, não em causalidade.
- Troque as variáveis de eixo, se ajudar a deixar isto claro.

4.4.1

Horas de sono e peso corporal

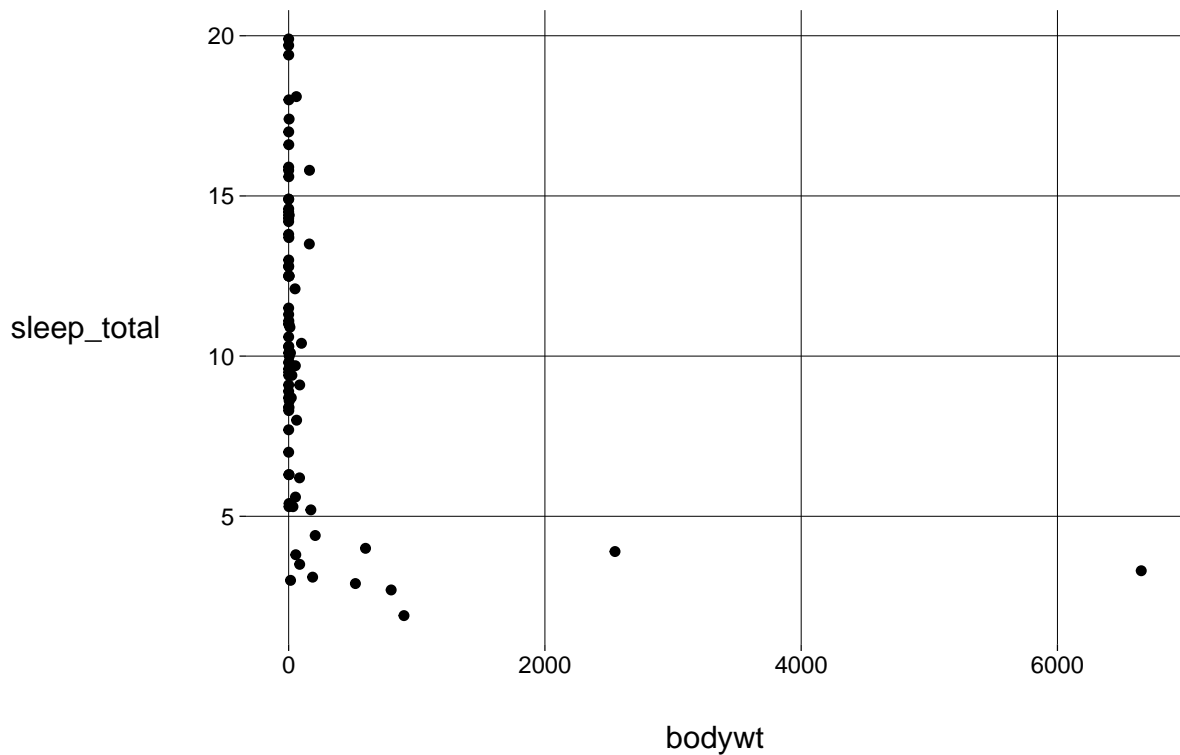
- Como as variáveis `sleep_total` e `bodywt` estão relacionadas?

```
sono %>%
  ggplot(aes(x = bodywt, y = sleep_total))
```



- O que houve? Cadê os pontos?
- O problema foi que só especificamos o mapeamento estético (com `aes`, que são as iniciais de *aesthetics*). **Faltou a geometria.**

```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
  geom_point()
```

- Que horror.
- A única coisa que percebemos aqui é que os mamíferos muito pesados dormem menos de 5 horas por noite.
- Estes animais muito pesados estão estragando a escala do eixo x .
- Que animais são estes?

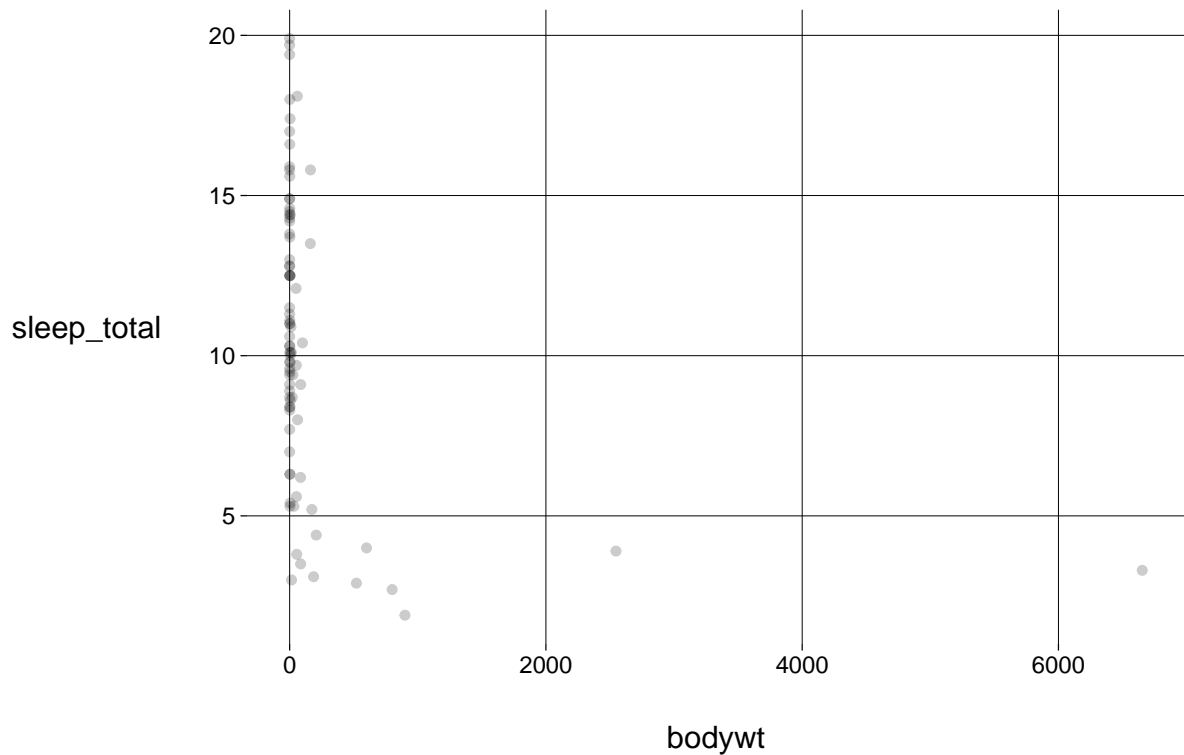
```
sono %>%
  filter(bodywt > 250) %>%
  select(name, bodywt) %>%
  arrange(bodywt)
```

```
## # A tibble: 6 x 2
##   name      bodywt
##   <chr>      <dbl>
## 1 Horse        521
## 2 Cow          600
## 3 Pilot whale  800
## 4 Giraffe     900.
## 5 Asian elephant 2547
## 6 African elephant 6654
```

- Além disso, há muitos pontos sobrepostos. Em bom português, temos um problema de *overplotting*.
- Existem diversas maneiras de lidar com isso.

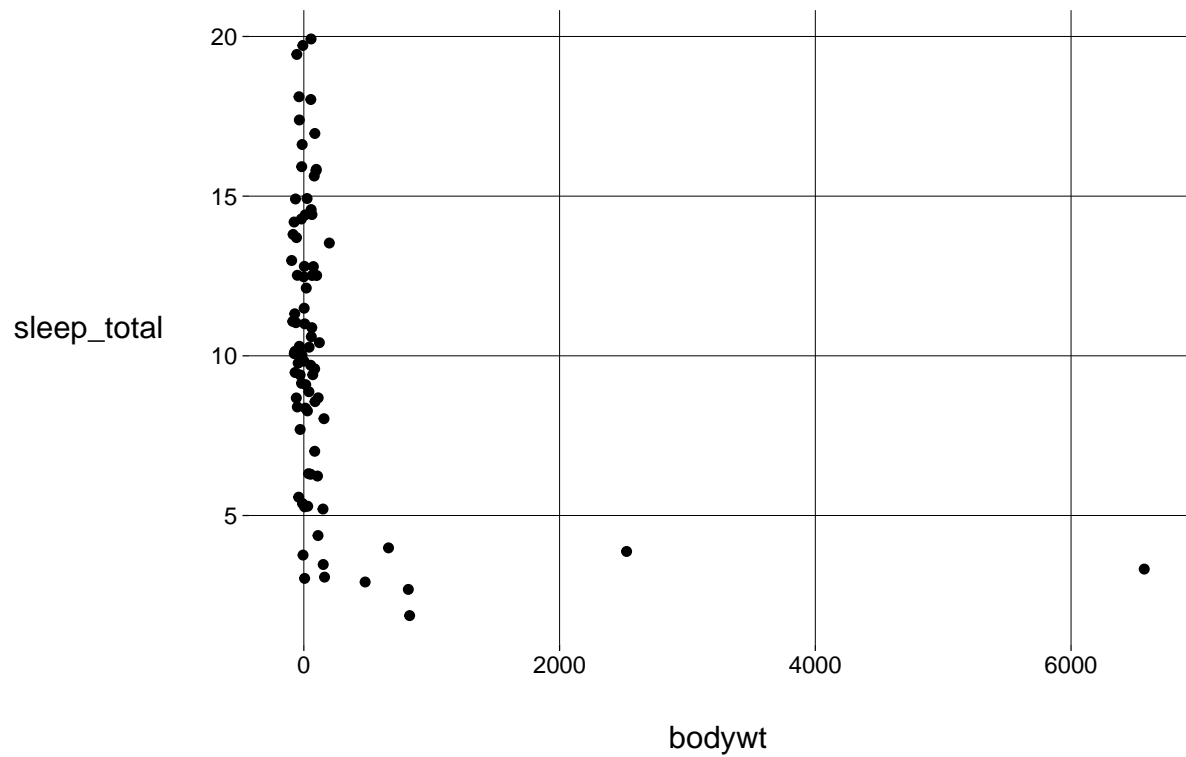
- A primeira delas é **alterando a opacidade dos pontos**. Isto é um ajuste na geometria apenas, pois a opacidade, aqui, não representa informação nenhuma.

```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
    geom_point(alpha = 0.2)
```



- Outra maneira é usar `geom_jitter` em vez de `geom_point`. “*Jitter*” significa “tremer”. As posições dos pontos são ligeiramente perturbadas, para evitar colisões. Perde-mos precisão, mas a visualização fica melhor.

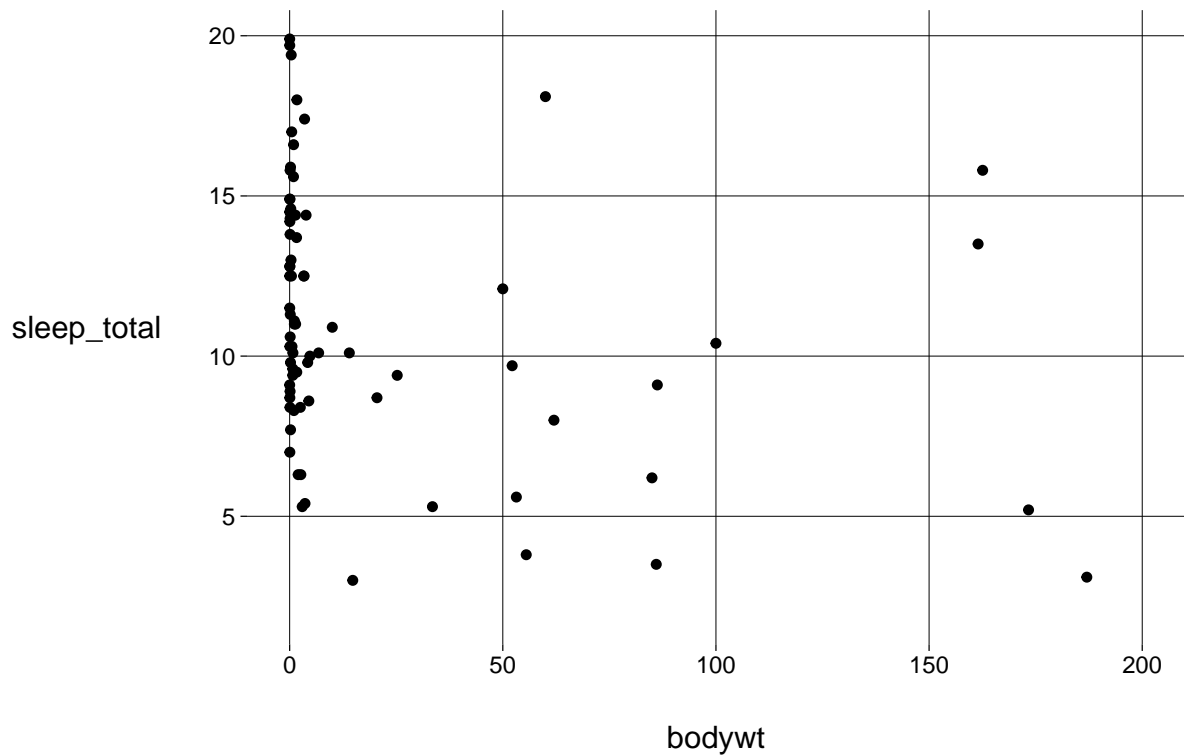
```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
    geom_jitter(width = 100)
```



- Vamos mudar os limites do gráfico para nos concentrarmos nos animais menos pesados. Observe que isto é um ajuste na escala.

```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
    geom_point() +  
    scale_x_continuous(limits = c(0, 200))
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



- Nestes limites, a relação entre horas de sono e peso não é mais tão pronunciada.

4.4.2

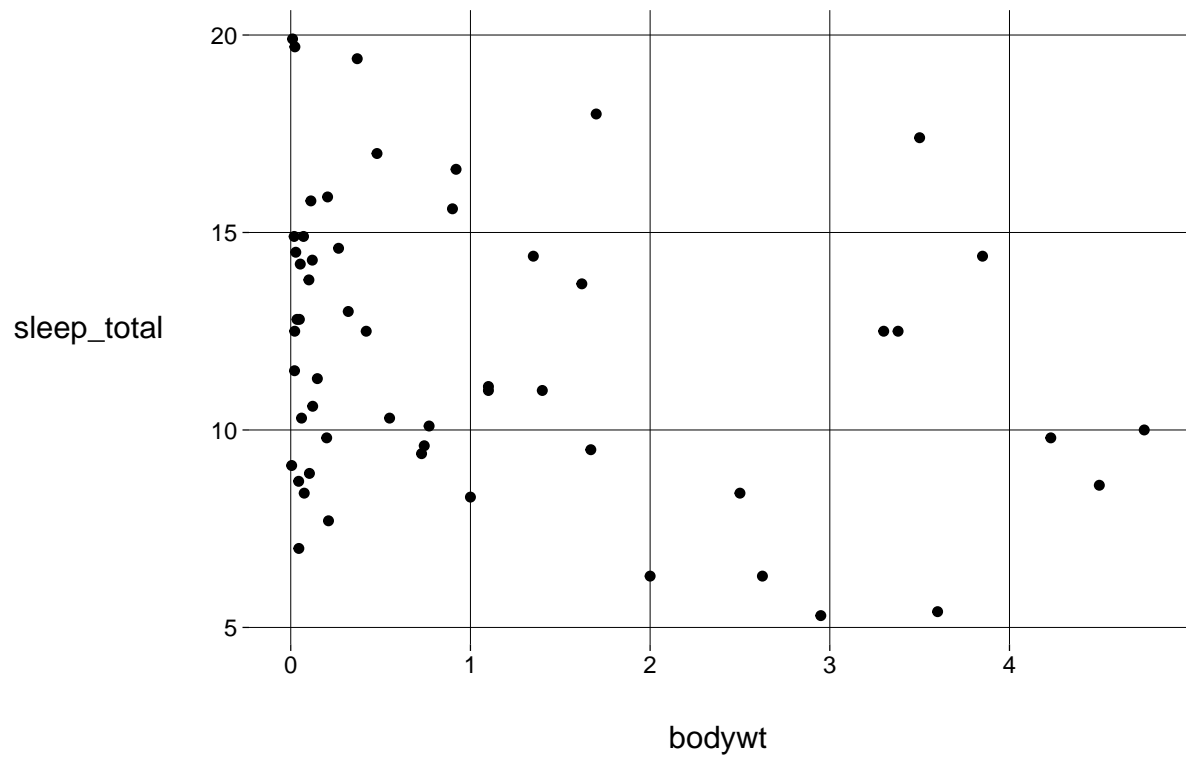
Horas de sono e peso corporal para animais pequenos

- Vamos restringir o gráfico a animais com no máximo 5kg.

```
limite <- 5
```

- Em vez de mudar a escala do gráfico, vamos filtrar as linhas do *data frame*:

```
sono %>%
  filter(bodywt < limite) %>%
  ggplot(aes(x = bodywt, y = sleep_total)) +
  geom_point()
```

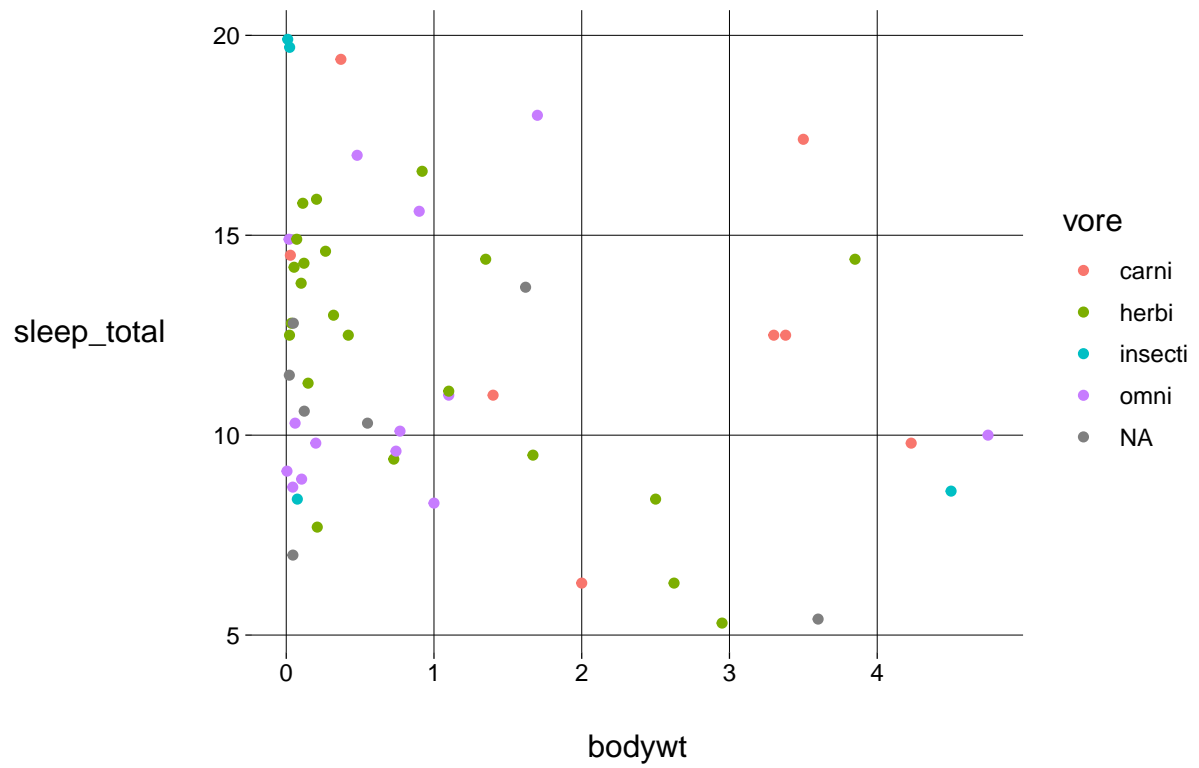


4.4.3

Incluindo a dieta

- Com a estética `color`. Observe como a legenda aparece automaticamente.

```
sono %>%  
  filter(bodywt < limite) %>%  
  ggplot(aes(x = bodywt, y = sleep_total, color = vore)) +  
    geom_point()
```

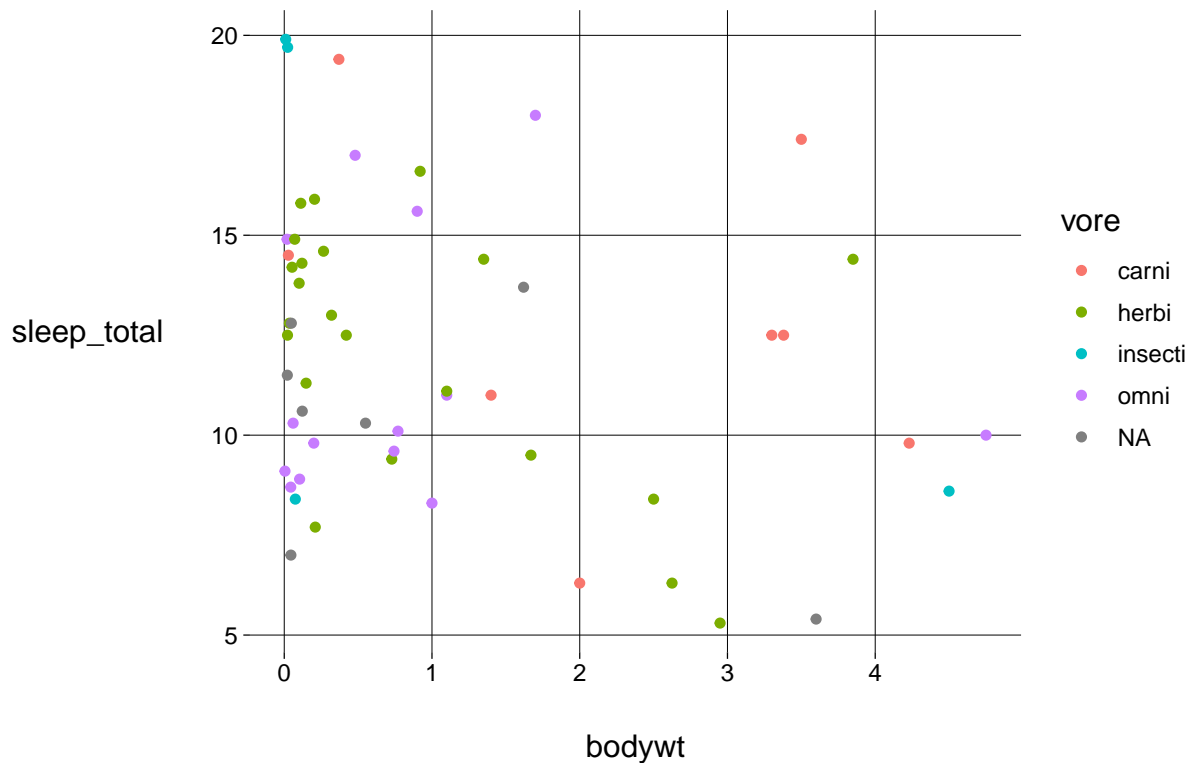


4.4.4

A estética pode ser especificada na `geom`

- Compare com o código anterior.

```
sono %>%
  filter(bodywt < limite) %>%
  ggplot() +
    geom_point(aes(x = bodywt, y = sleep_total, color = vore))
```



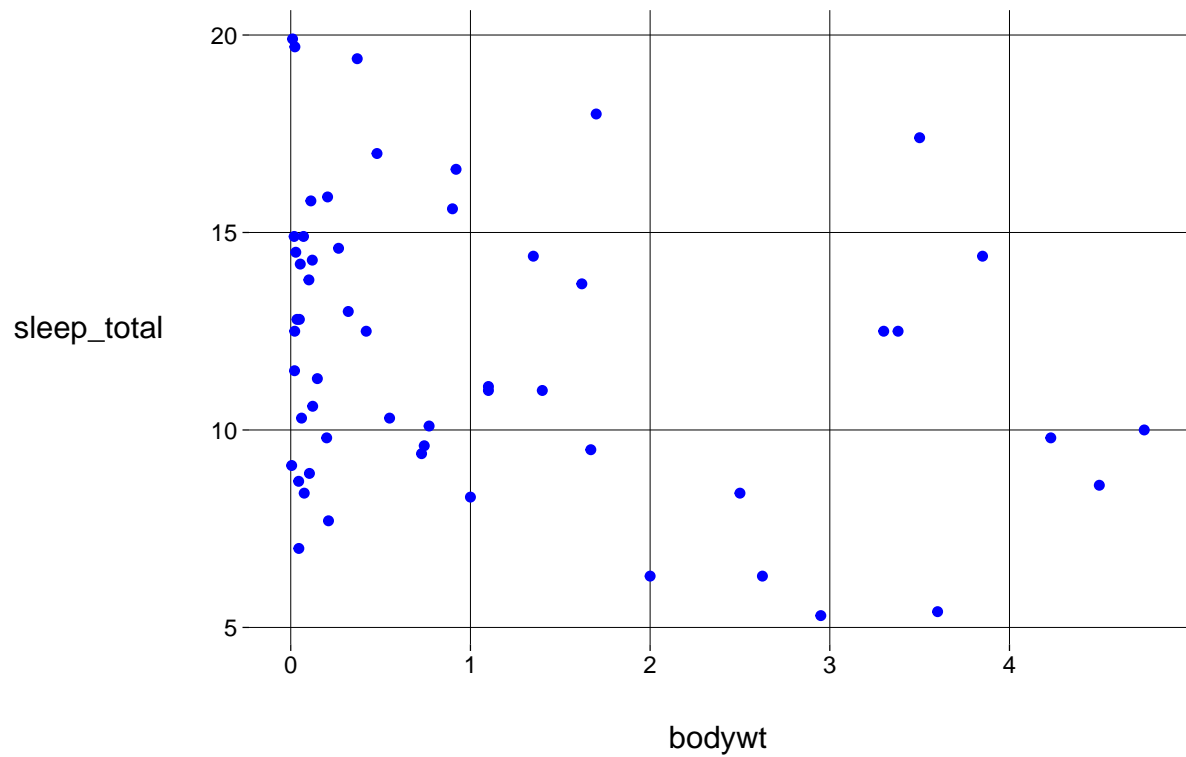
- Fazendo deste modo, a estética só vale para uma geometria. Se você acrescentar outras geometrias (linhas, por exemplo), a estética não valerá para elas.

4.4.5

Aparência fixa ou dependendo de variável?

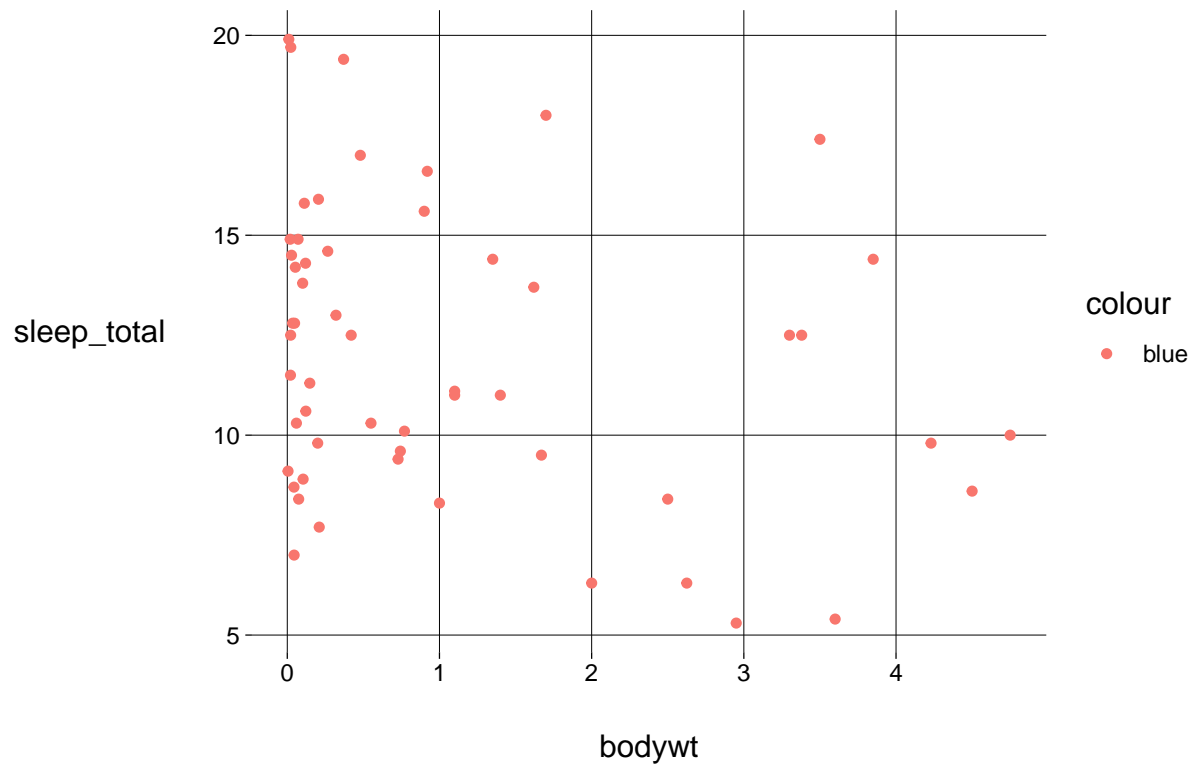
- Se for fixa, não é estética. Não representa informação.
- Se depender de variável, é estética. Representa informação.
- Compare o último *chunk* acima com:

```
sono %>%
  filter(bodywt < limite) %>%
  ggplot() +
    geom_point(aes(x = bodywt, y = sleep_total), color = 'blue')
```



- Se for uma estética, precisa estar associada a uma variável, não a um valor fixo. Um erro comum seria fazer:

```
sono %>%  
  filter(bodywt < limite) %>%  
  ggplot() +  
    geom_point(aes(x = bodywt, y = sleep_total, color = 'blue'))
```

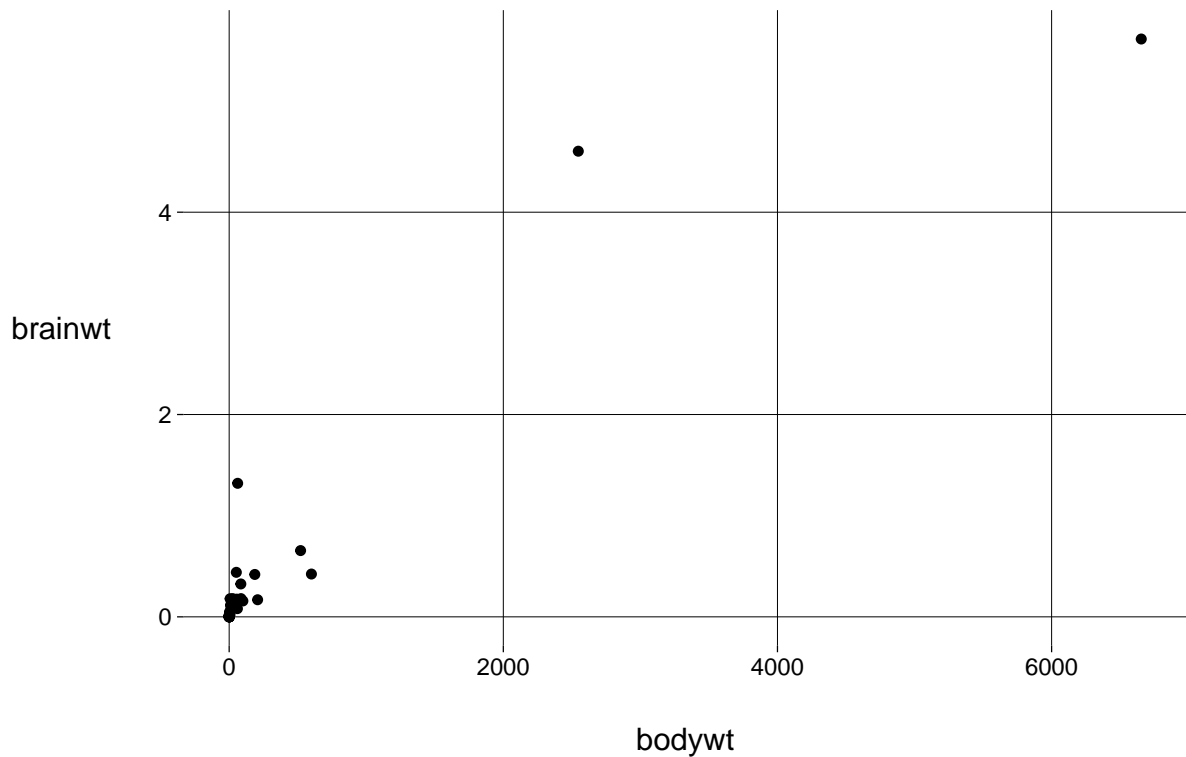
4.4.6

Uma correlação mais clara

- Peso cerebral versus peso corporal:

```
sono %>%  
  ggplot(aes(x = bodywt, y = brainwt)) +  
    geom_point()
```

```
## Warning: Removed 27 rows containing missing values (geom_point).
```



- A mensagem de aviso (*warning*) diz que há 27 valores faltantes (NA) em `bodywt` ou `brainwt`. De fato:

```
sono %>%
  filter(is.na(bodywt)) %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```
sono %>%
  filter(is.na(brainwt)) %>%
  count()
```

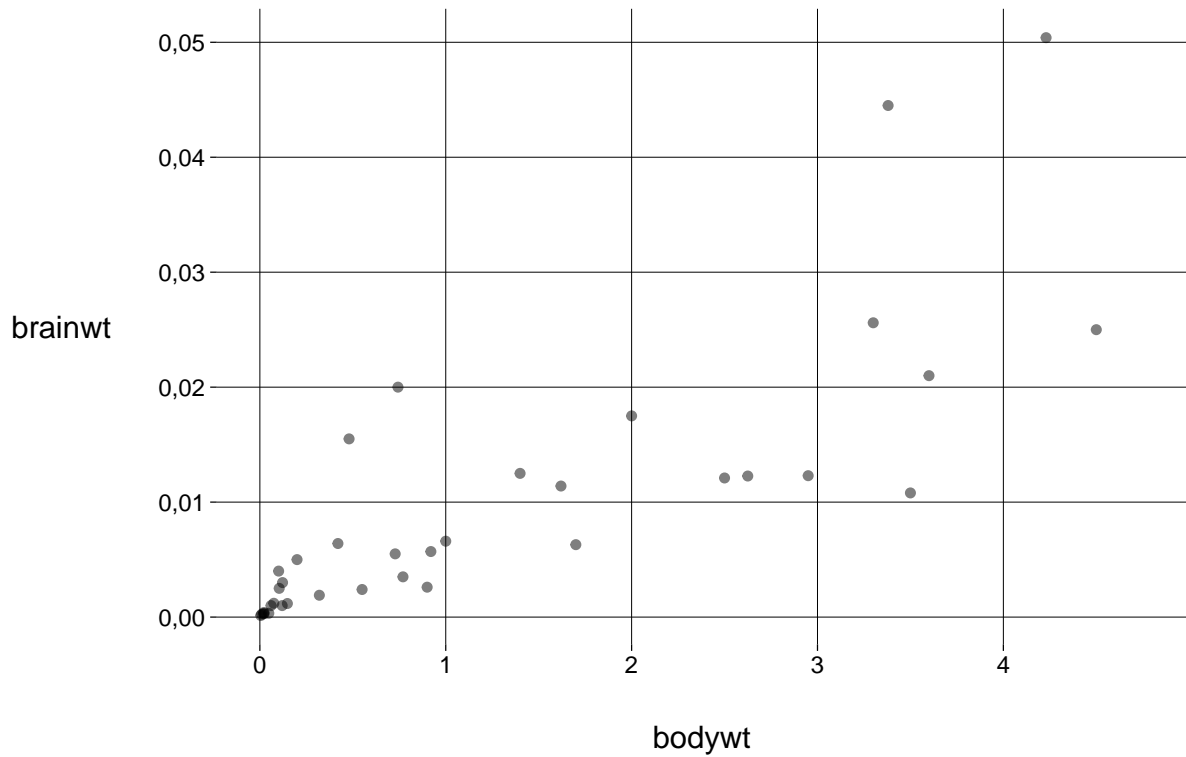
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    27
```

- Vamos restringir aos animais mais leves e mudar a opacidade:

```
sono %>%
  filter(bodywt < limite) %>%
  ggplot(aes(x = bodywt, y = brainwt)) +
```

```
geom_point(alpha = .5)
```

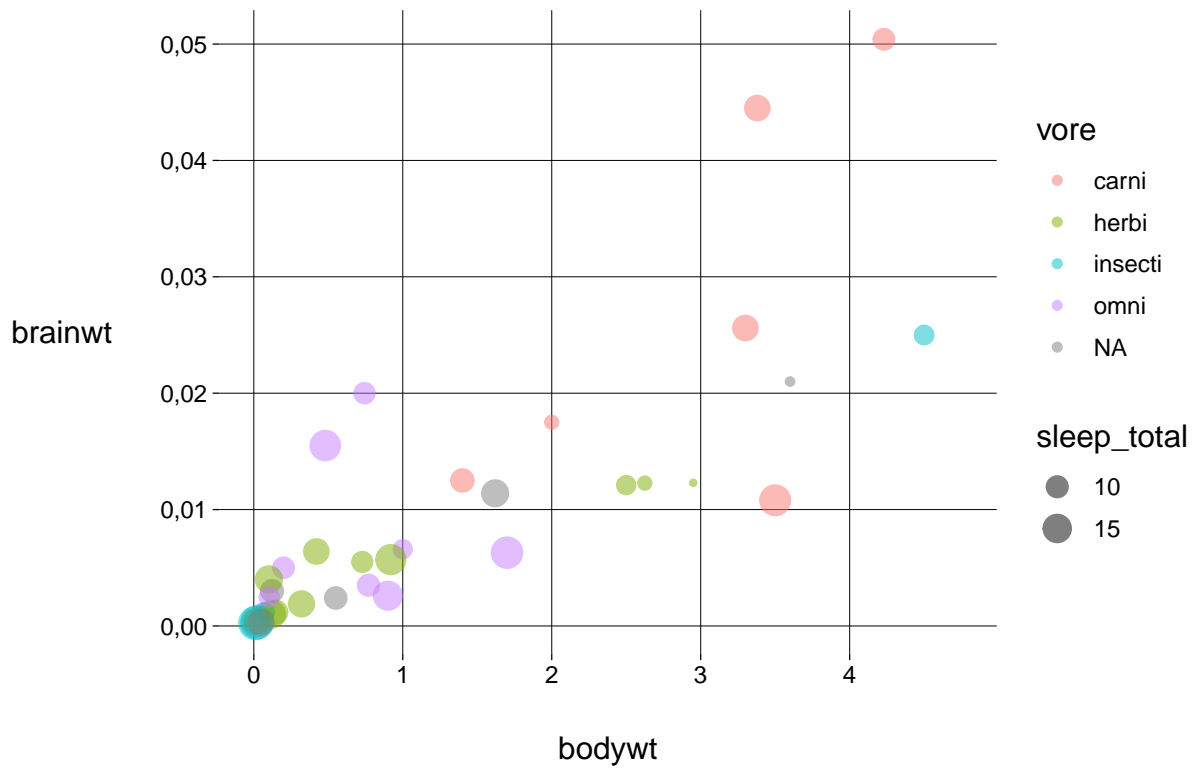
```
## Warning: Removed 18 rows containing missing values (geom_point).
```



- Vamos incluir horas de sono e dieta. Observe as estéticas usadas.

```
sono %>%  
  filter(bodywt < limite) %>%  
  ggplot(  
    aes(  
      x = bodywt,  
      y = brainwt,  
      size = sleep_total,  
      color = vore  
    )  
  ) +  
  geom_point(alpha = .5)
```

```
## Warning: Removed 18 rows containing missing values (geom_point).
```



- Vamos mudar a escala dos tamanhos e incluir rótulos:

```
grafico <- sono %>%
  filter(bodywt < limite) %>%
  ggplot(
    aes(
      x = bodywt,
      y = brainwt,
      size = sleep_total,
      color = vore
    )
  ) +
  geom_point(alpha = .5) +
  scale_size(
    breaks = seq(0, 24, 4)
  ) +
  labs(
    title = 'Peso do cérebro versus peso corporal',
    subtitle = paste0(
      'para mamíferos com menos de ',
      limite,
      ' kg'
    ),
    caption = 'Fonte: dataset `msleep`',
    x = 'Peso corporal (kg)',
  )
```

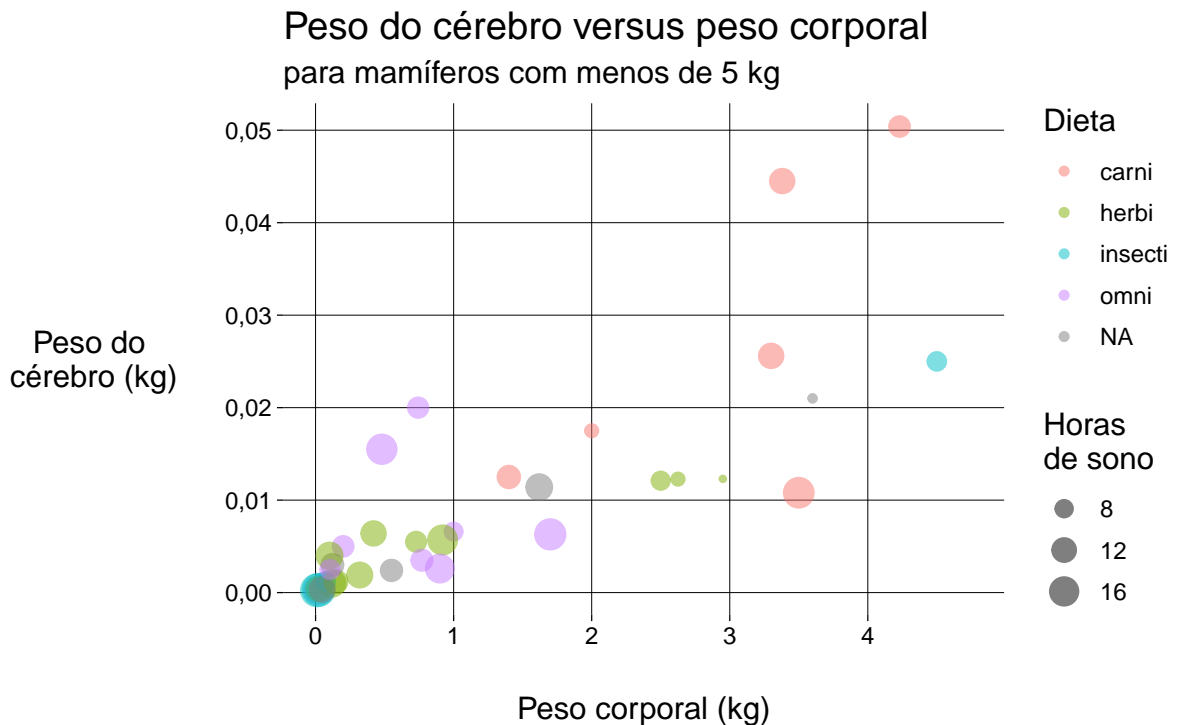
```

y = 'Peso do\n cérebro (kg)',
color = 'Dieta',
size = 'Horas\nde sono'
)

```

grafico

Warning: Removed 18 rows containing missing values (geom_point).



- Vamos mudar as cores usadas para a dieta, usando uma escala diferente.

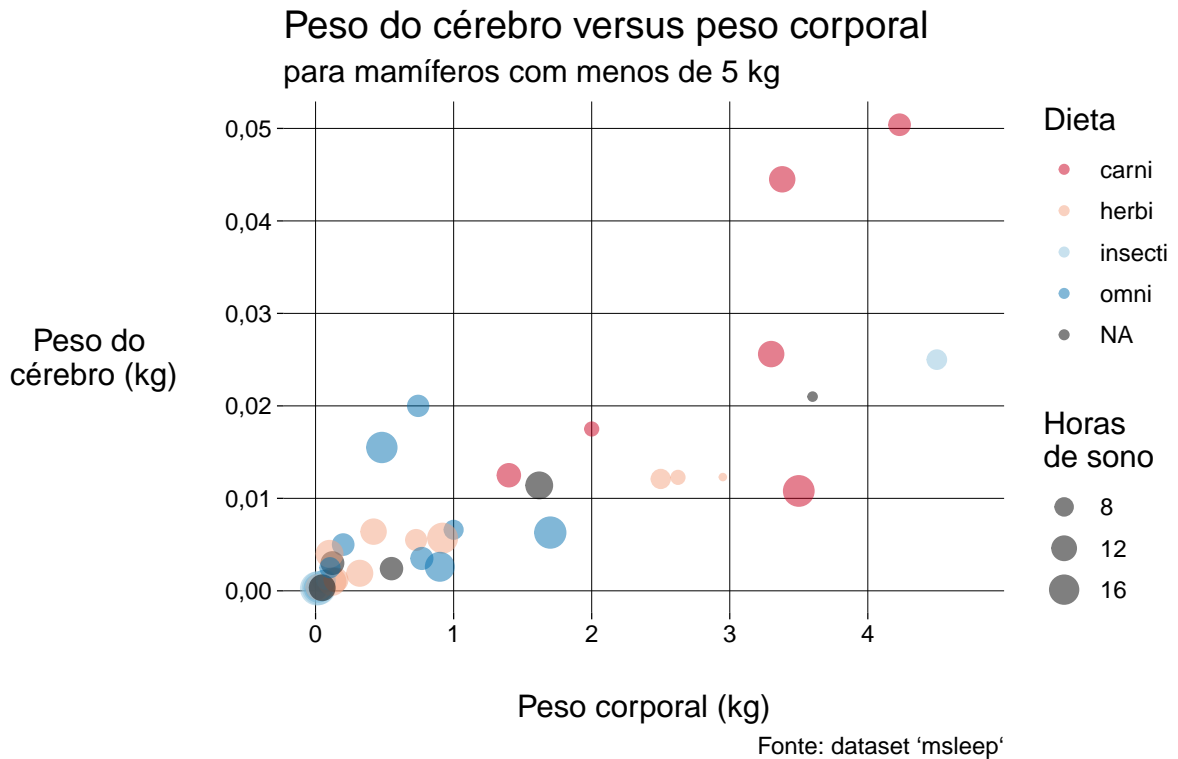
```

grafico2 <- grafico +
  scale_color_discrete(
    palette = 'RdBu',
    na.value = 'black',
    type = scale_color_brewer
  )

```

grafico2

Warning: Removed 18 rows containing missing values (geom_point).



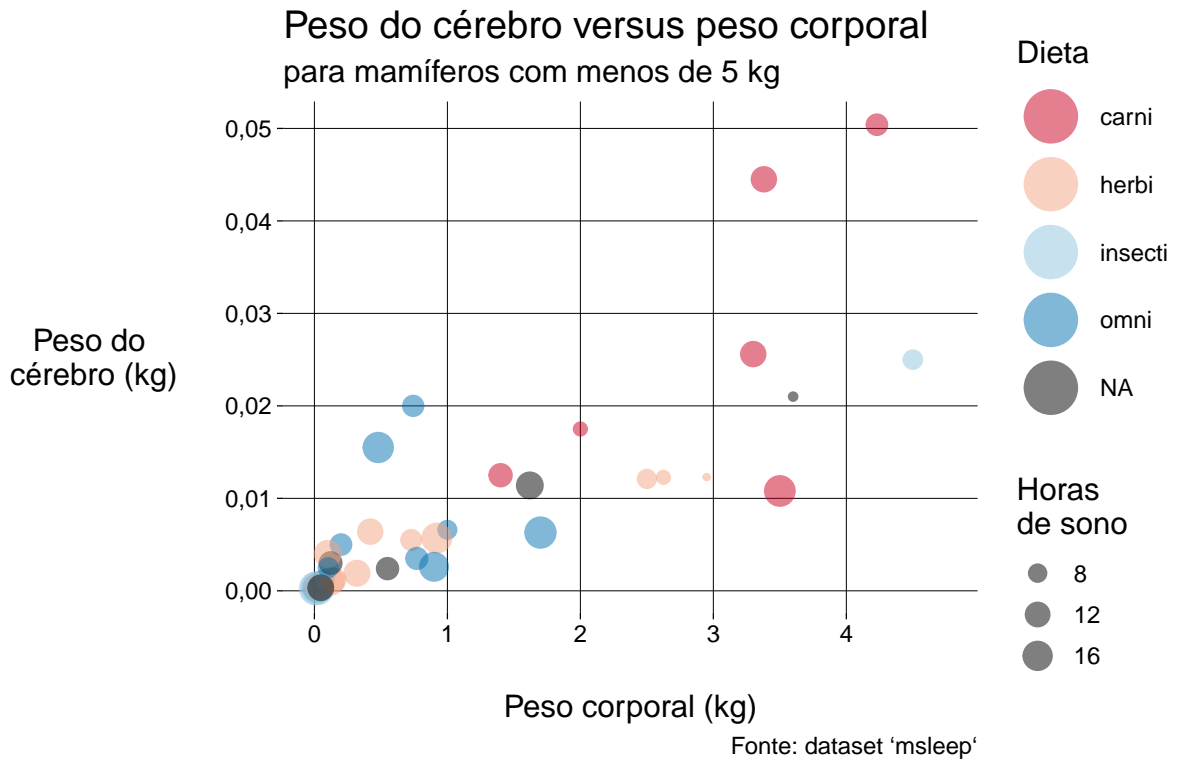
- Observe como usamos o gráfico já salvo na variável `grafico` e simplesmente acrescentamos a nova escala. Este tipo de “montagem” de gráficos `ggplot2` é bem conveniente, para evitar repetição de código.
- Um último ajuste na aparência: os pontos na legenda “Dieta” estão pequenos demais. Quase não identificamos as cores deles.

Vamos usar a função `guides` para modificar (*override*) a estética `color` — apenas na legenda, não nos pontos mostrados no gráfico, cujos tamanhos representam o número de horas de sono — tornando o tamanho maior. [Leia mais sobre `override.aes` neste link \(em inglês\)](#).

```
grafico3 <- grafico2 +
  guides(color = guide_legend(override.aes = list(size = 10)))

grafico3
```

Warning: Removed 18 rows containing missing values (geom_point).



- Agora podemos finalmente comentar sobre a informação que o gráfico mostra sobre os dados:
 - De fato, existe uma correlação entre peso cerebral e peso corporal: quanto maior o peso corporal, maior o peso cerebral. Nada surpreendente.
 - Podemos fazer o `ggplot2` traçar uma reta de regressão com a geometria `geom_smooth`. Vamos falar mais sobre correlação **em um capítulo futuro**.

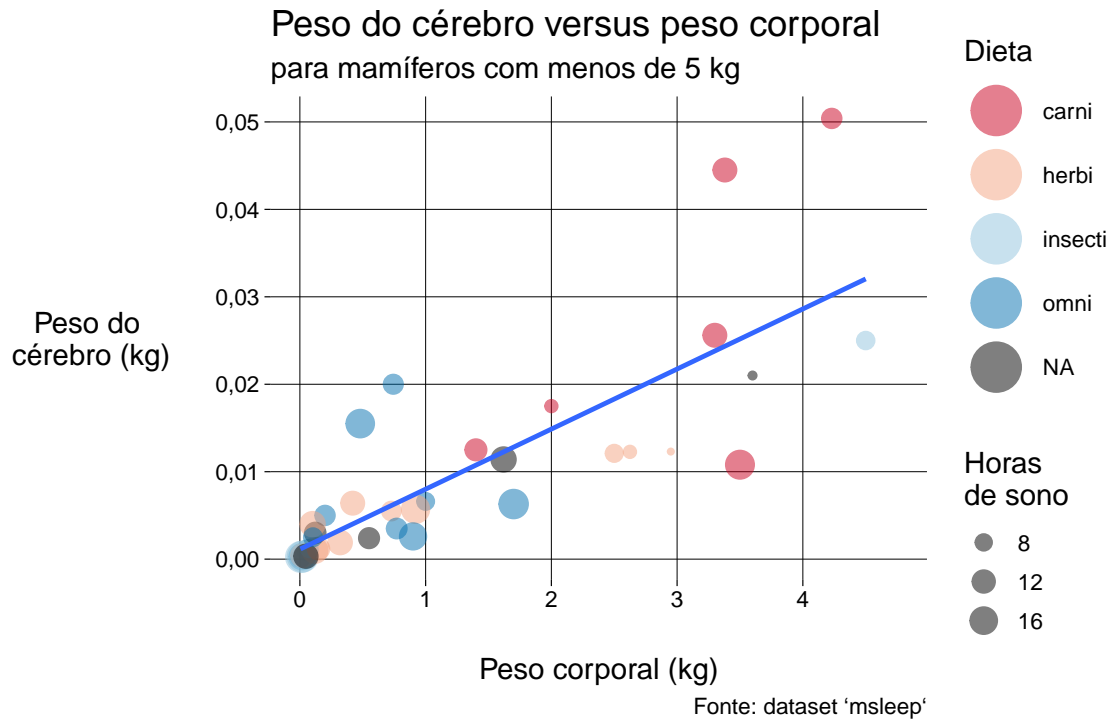
```
grafico4 <- grafico3 +
  geom_smooth(
    aes(group = 1),
    show.legend = FALSE,
    method = 'lm',
    se = FALSE
  )
```

```
grafico4
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 18 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 18 rows containing missing values (geom_point).
```



- Todos os carnívoros têm peso corporal maior que 1kg e peso cerebral maior ou igual a 10g.
- Só um carnívoro dorme 8 horas ou menos. Qual?
- Todos os insetívoros — com exceção de um (qual?) — são muito leves e dormem muito.
- Todos os onívoros têm menos de 2kg de peso corporal e 20g ou menos de peso cerebral.

4.5

Vídeo 2

<https://youtu.be/c-LoZ9e8xWc>

4.6

Histogramas e cia.

- A idéia agora é **agrupar indivíduos em classes**, dependendo do valor de uma variável quantitativa.

4.6.1

Distribuições de frequência

- Vamos nos concentrar nas horas de sono.

```
sono$sleep_total
```

```
## [1] 12,1 17,0 14,4 14,9 4,0 14,4 8,7 7,0 10,1 3,0 5,3 9,4 10,0
## [14] 12,5 10,3 8,3 9,1 17,4 5,3 18,0 3,9 19,7 2,9 3,1 10,1 10,9
## [27] 14,9 12,5 9,8 1,9 2,7 6,2 6,3 8,0 9,5 3,3 19,4 10,1 14,2
## [40] 14,3 12,8 12,5 19,9 14,6 11,0 7,7 14,5 8,4 3,8 9,7 15,8 10,4
## [53] 13,5 9,4 10,3 11,0 11,5 13,7 3,5 5,6 11,1 18,1 5,4 13,0 8,7
## [66] 9,6 8,4 11,3 10,6 16,6 13,8 15,9 12,8 9,1 8,6 15,8 4,4 15,6
## [79] 8,9 5,2 6,3 12,5 9,8
```

- Antes de montar o histograma, vamos construir uma **distribuição de frequência**.
- A **amplitude** é a diferença entre o valor máximo e o valor mínimo. A função `range` não retorna a amplitude, mas sim os valores mínimo e máximo:

```
sono$sleep_total %>% range()
```

```
## [1] 1,9 19,9
```

- Vamos decidir que cada classe vai ter 2 horas. A função `cut` substitui os valores do vetor pelos nomes das classes:

```
sono$sleep_total %>%
  cut(breaks = seq(0, 20, 2), right = FALSE)
```

```
## [1] [12,14) [16,18) [14,16) [14,16) [4,6) [14,16) [8,10) [6,8)
## [9] [10,12) [2,4) [4,6) [8,10) [10,12) [12,14) [10,12) [8,10)
## [17] [8,10) [16,18) [4,6) [18,20) [2,4) [18,20) [2,4) [2,4)
## [25] [10,12) [10,12) [14,16) [12,14) [8,10) [0,2) [2,4) [6,8)
## [33] [6,8) [8,10) [8,10) [2,4) [18,20) [10,12) [14,16) [14,16)
## [41] [12,14) [12,14) [18,20) [14,16) [10,12) [6,8) [14,16) [8,10)
## [49] [2,4) [8,10) [14,16) [10,12) [12,14) [8,10) [10,12) [10,12)
## [57] [10,12) [12,14) [2,4) [4,6) [10,12) [18,20) [4,6) [12,14)
## [65] [8,10) [8,10) [8,10) [10,12) [10,12) [16,18) [12,14) [14,16)
## [73] [12,14) [8,10) [8,10) [14,16) [4,6) [14,16) [8,10) [4,6)
## [81] [6,8) [12,14) [8,10)
## 10 Levels: [0,2) [2,4) [4,6) [6,8) [8,10) [10,12) [12,14) ... [18,20)
```

- A função `table` faz a contagem dos elementos de cada classe:

```
sono$sleep_total %>%
  cut(breaks = seq(0, 20, 2), right = FALSE) %>%
  table(dnn = 'Horas de sono') %>%
```

```
as.data.frame()

## # A tibble: 10 x 2
##   Horas.de.sono Freq
##   <fct>         <int>
## 1 [0,2)         1
## 2 [2,4)         8
## 3 [4,6)         7
## 4 [6,8)         5
## 5 [8,10)        17
## 6 [10,12)       14
## # ... with 4 more rows
```

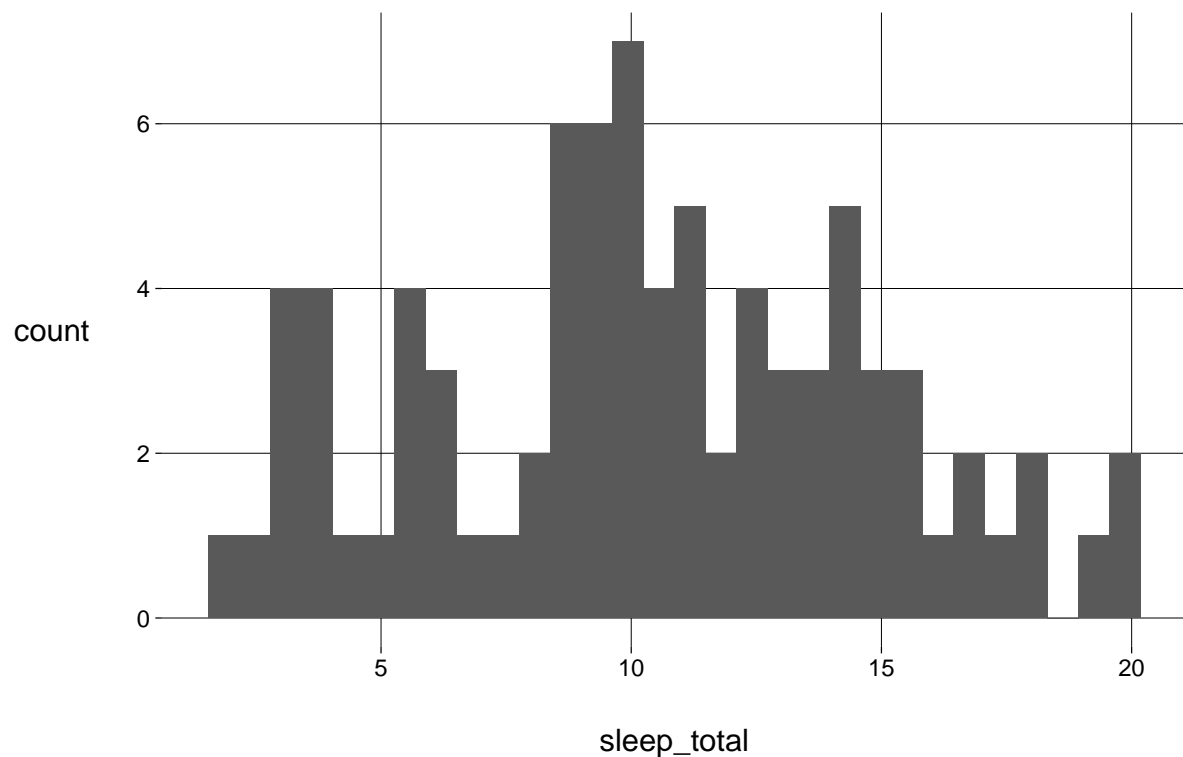
4.6.2

Histograma

- Na verdade, o `ggplot2` já faz esses cálculos para nós.
- O *default* é criar 30 classes (*bins*):

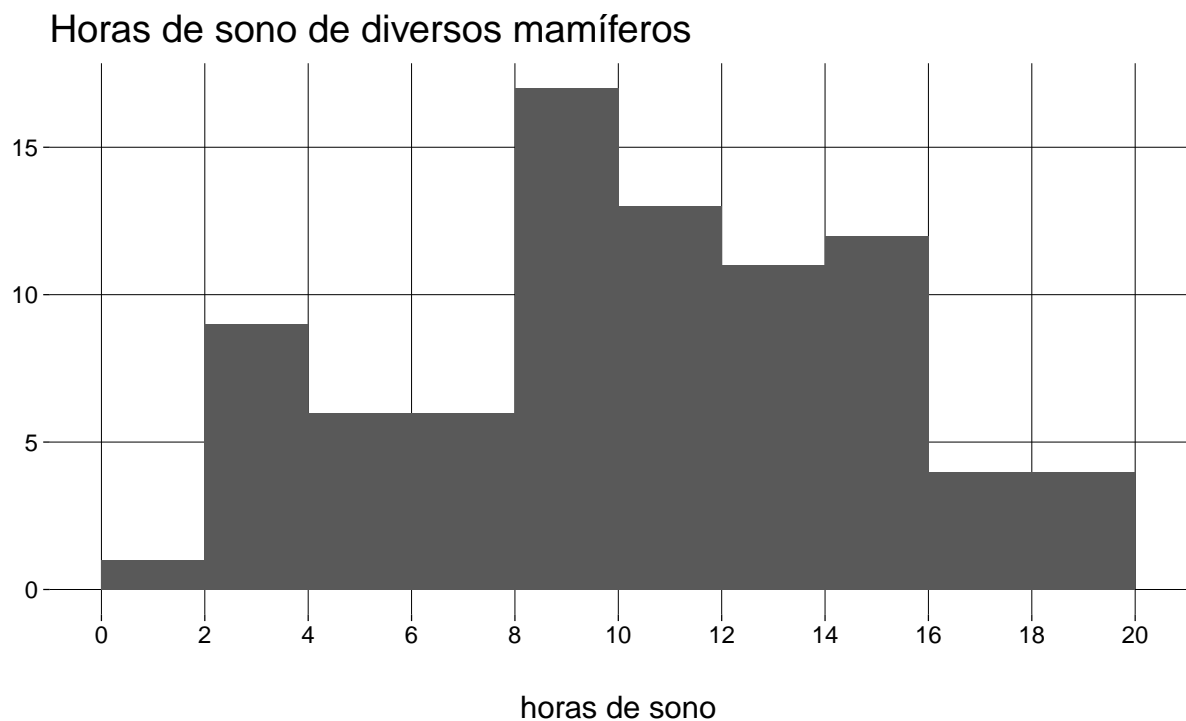
```
sono %>%
  ggplot(aes(x = sleep_total)) +
    geom_histogram()
```

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



- Vamos mudar isto passando um vetor de limites das classes (*breaks*). Vamos acrescentar rótulos também:

```
sono %>%
  ggplot(aes(x = sleep_total)) +
  geom_histogram(breaks = seq(0, 20, 2)) +
  scale_x_continuous(breaks = seq(0, 20, 2)) +
  labs(
    title = 'Horas de sono de diversos mamíferos',
    x = 'horas de sono',
    y = NULL,
    caption = 'Fonte: dataset `msleep`'
  )
```



Fonte: dataset 'msleep'

- Nossas impressões:
 - A classe que mais tem elementos é a de 8 a 10 horas.
 - A distribuição é mais ou menos simétrica.
 - A distribuição tem forma aproximada de sino: há poucos mamíferos com valores extremos de horas de sono; a maioria está próxima do valor médio:

```
mean(sono$sleep_total)
```

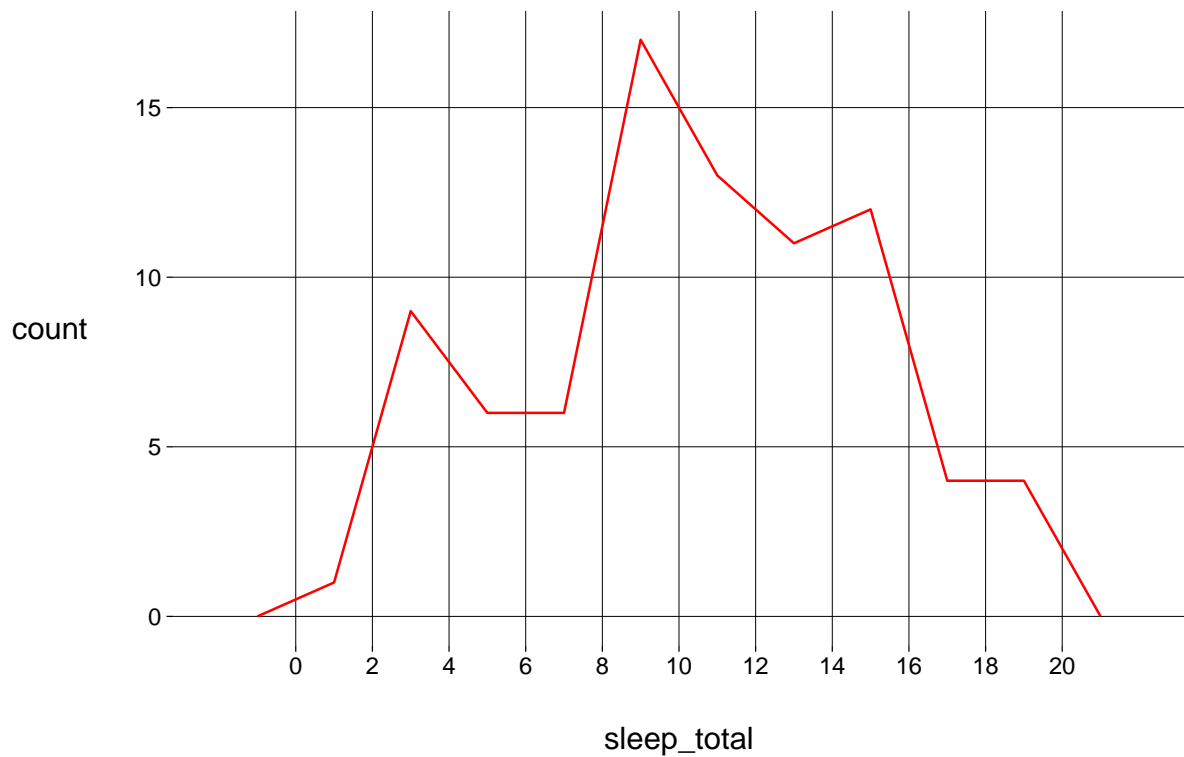
```
## [1] 10,43373
```

4.6.3

Polígono de frequência

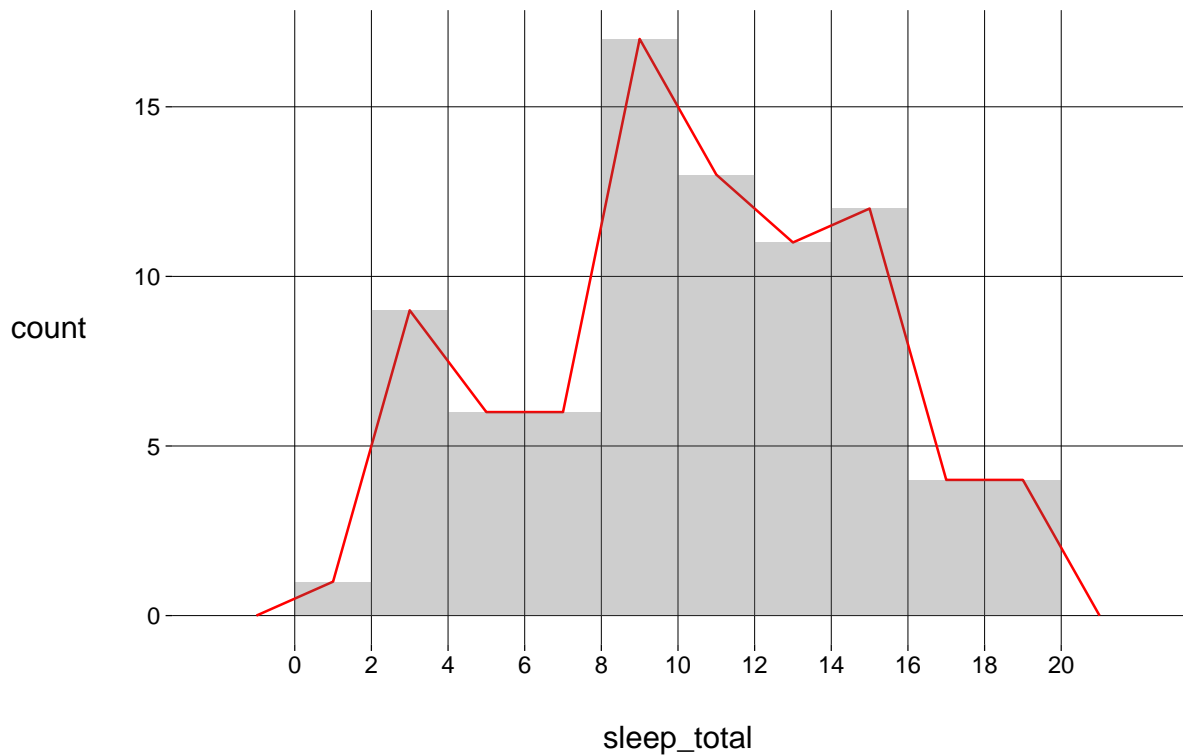
- Em vez das barras do histograma, podemos desenhar uma linha ligando seus topos.
- O resultado é um **polígono de frequência**.

```
pf <- sono %>%  
  ggplot(aes(x = sleep_total)) +  
    geom_freqpoly(breaks = seq(0, 20, 2), color = 'red') +  
    scale_x_continuous(breaks = seq(0, 20, 2))  
  
pf
```



- Vamos sobrepor o polígono de frequência ao histograma, para deixar claro o que está acontecendo:

```
pf + geom_histogram(breaks = seq(0, 20, 2), alpha = .3)
```

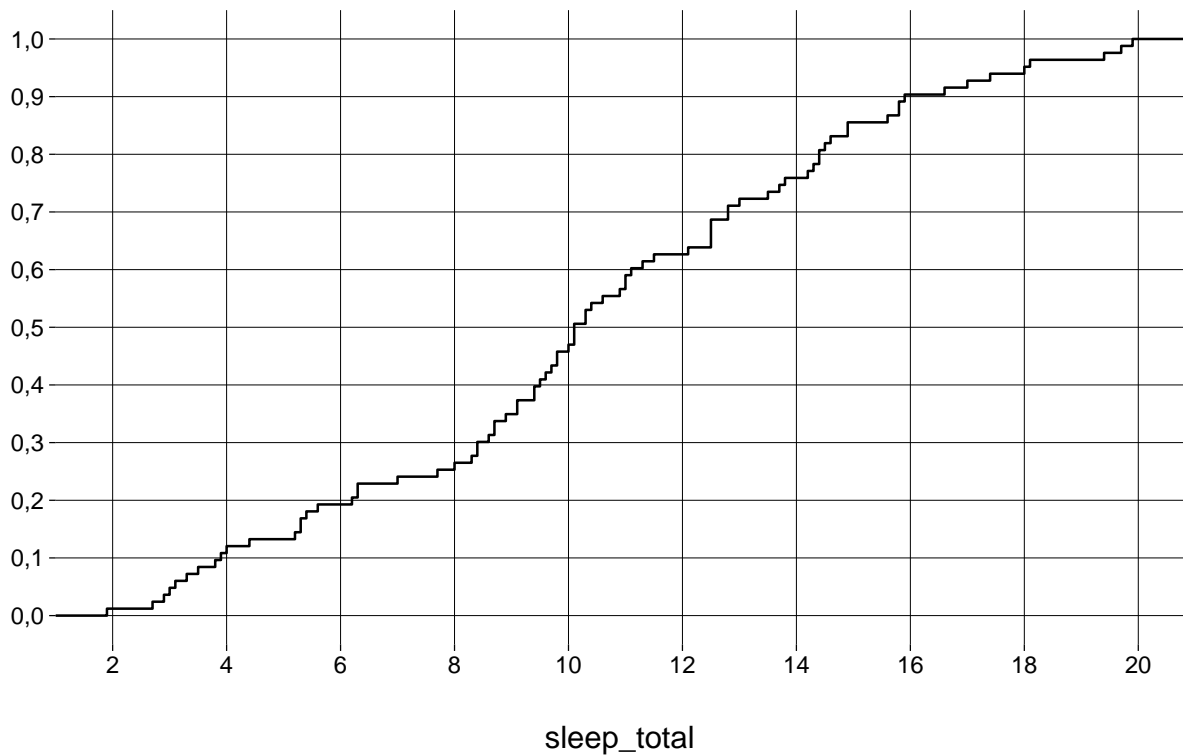


4.7

Ogiva

- A ogiva é um gráfico que mostra a **frequência acumulada**: para cada valor v da variável no eixo x , a proporção de indivíduos com valor menor ou igual a v .
- A geometria `geom_step` gera o gráfico de uma **função degrau**.
- Cada geometria está ligada a uma **stat**, um algoritmo para computar o que vai ser desenhado. Aqui, passamos para a geometria a função **`ecdf` (empirical cumulative distribution function)**, do pacote `stats`, que calcula as frequências acumuladas.

```
sono %>%
  ggplot(aes(x = sleep_total)) +
    geom_step(stat = 'ecdf') +
    scale_x_continuous(breaks = seq(0, 20, 2)) +
    scale_y_continuous(breaks = seq(0, 1, .1)) +
    labs(y = NULL)
```



- Com a ogiva, podemos obter informações difíceis de visualizar no histograma. Por exemplo:
 - Cerca de 20% dos mamíferos têm menos de 6 horas de sono.
 - Cerca de metade dos mamíferos têm menos de 10 horas de sono.
 - Cerca de 10% dos mamíferos têm mais de 16 horas de sono.

4.8

Ramos e folhas

- No início dos anos 1900, quando estatísticas eram feitas à mão, Arthur Bowley criou os **diagramas de ramos e folhas**.
- Um diagrama de ramos e folhas é, basicamente, uma listagem de todos os valores de uma variável, agrupados de maneira que todos os valores de uma classe (i.e., de uma linha) têm os algarismos iniciais dentro de um intervalo.
- Para as horas de sono dos mamíferos:

```
sono$sleep_total %>%
  stem()
```

```
##
## The decimal point is at the |
##
```

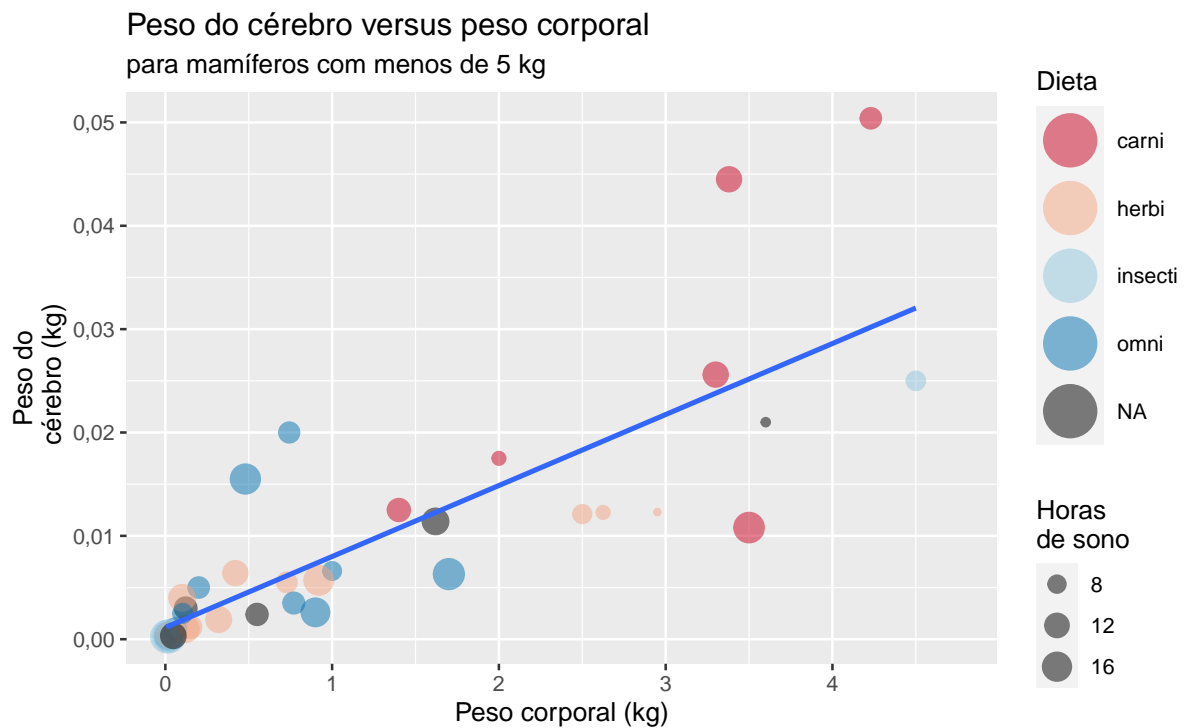
```
##    0 | 9
##    2 | 79013589
##    4 | 0423346
##    6 | 23307
##    8 | 03446779114456788
##   10 | 01113346900135
##   12 | 15555880578
##   14 | 234456996889
##   16 | 604
##   18 | 01479
```

- A primeira linha representa um indivíduo com 0,9 horas de sono.
- A penúltima linha representa 3 valores:
 - 16,6
 - 17,0
 - 17,4

4.9

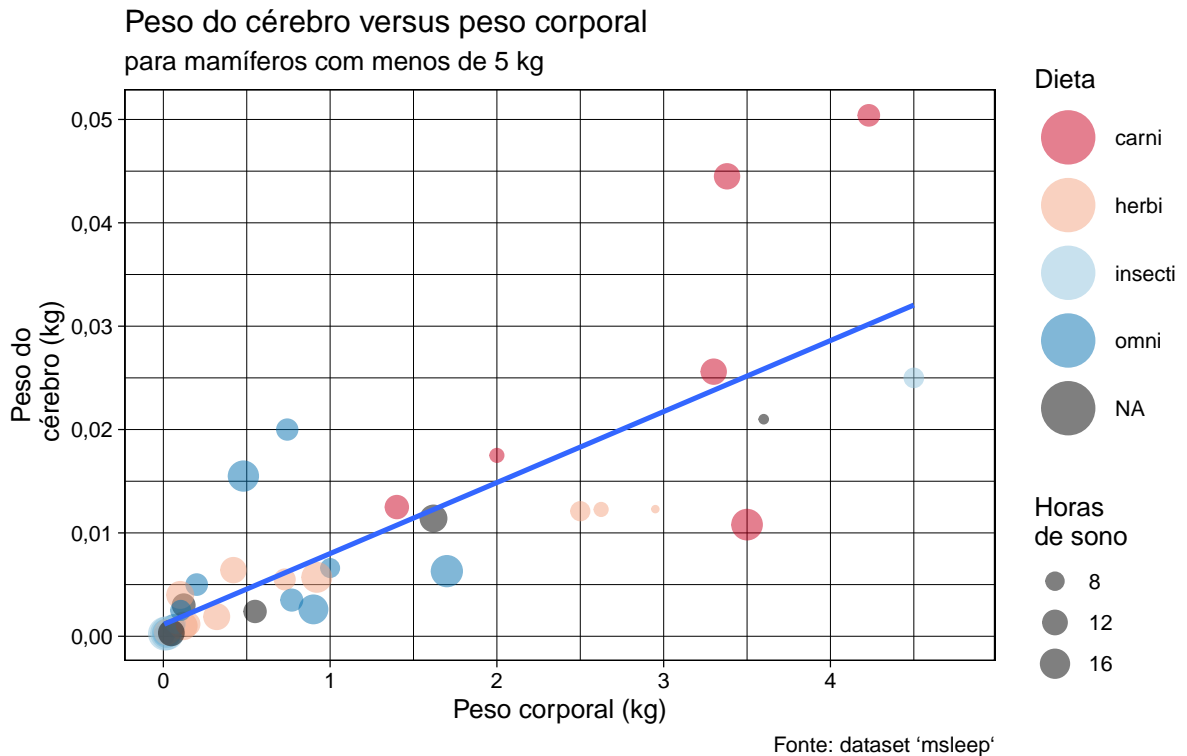
Personalização do tema

- O `ggplot2` tem um tema *default*, chamado `theme_gray`, que gera o *scatterplot* de um exemplo anterior deste capítulo do seguinte modo:

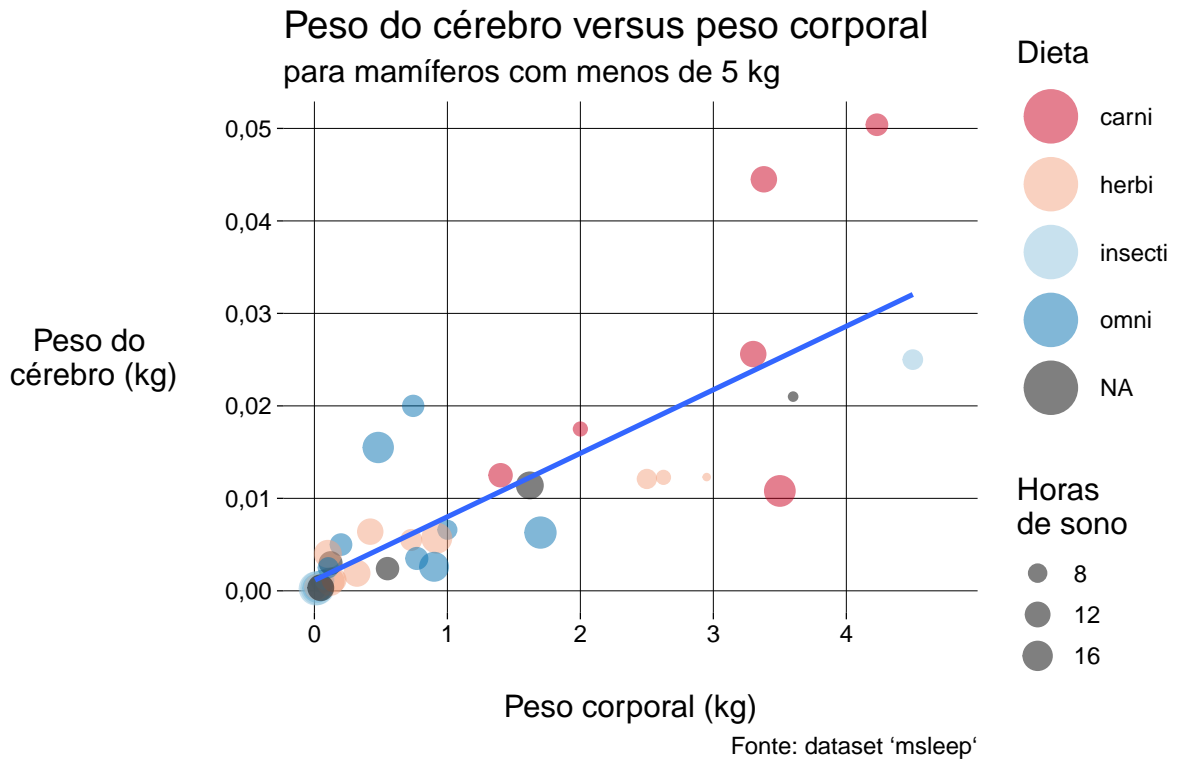


- Para este material, escolhi o tema `theme_linedraw`, que usa linhas pretas sobre fundo

branco:



- Para deixar os gráficos mais leves e facilitar a leitura, fiz as seguintes alterações no tema:
 - Mudei o tamanho do texto dos rótulos.
 - Fiz o rótulo do eixo y aparecer na horizontal; embora isto ocupe um pouco mais de espaço, evita que o leitor tenha que girar a cabeça para ler o rótulo.
 - Eliminei as linhas dos eixos, para o gráfico ficar mais leve.
 - Eliminei a moldura da área de dados, para o gráfico ficar mais leve.
 - Eliminei a grade secundária, para o gráfico ficar mais leve.
- O resultado é



- Os meus comandos para alterar o tema são

```
# Tamanho do texto depende do formato de saída (html ou pdf):
plot_text_size = ifelse(is_html_output(), 12, 13)

# Tema mais leve:
theme_set(
  theme_linedraw() +
  theme(
    # Tamanho do texto
    text = element_text(size = plot_text_size),
    # Eixo y
    axis.title.y.left = element_text(
      # Nunca girar o rótulo do eixo y
      angle = 0,
      # Separar o rótulo do eixo um pouco
      margin = margin(r = 20),
      # Posicionar verticalmente no meio
      vjust = .5
    ),
    # Eixo y secundário (à direita), quando presente
    axis.title.y.right = element_text(
      # Nunca girar o rótulo do eixo y
      angle = 0,
      # Separar o rótulo do eixo um pouco
```

```

margin = margin(l = 20),
# Posicionar verticalmente no meio
vjust = .5
),
# Não colocar marcas no eixo y secundário
axis.ticks.y.right = element_blank(),
# Separar o eixo x do rótulo um pouco mais
axis.title.x.bottom = element_text(
  margin = margin(t = 20)
),
# Eliminar linhas dos eixos
axis.line = element_blank(),
# Eliminar a moldura da área de dados
panel.border = element_blank(),
# Eliminar a grade secundária
panel.grid.minor = element_blank()
)
)

```

4.10

Exercícios



Não se esqueça de incluir títulos nos gráficos e rótulos nos eixos.

4.10.1

Peso cerebral e peso corporal

1. Observe os **comandos que geraram o gráfico grafico4**.
2. O que acontece se você retirar `aes(group = 1)` da chamada a `geom_smooth`? Explique.
3. O que acontece se você mudar `show.legend = FALSE` para `show.legend = TRUE` na chamada a `geom_smooth`? Explique.
4. O que acontece se você mudar `se = FALSE` para `se = TRUE` na chamada a `geom_smooth`? Explique.
5. Acrescente ao gráfico a camada `facet_wrap(~vore)`. O que acontece?
6. Examine o *data frame* `sono` e identifique o único insetívoro com mais de 4kg.
7. Instale o pacote `gg_repel` e acrescente ao gráfico `grafico4` (não facetado) a geometria `geom_label_repel` (consulte a ajuda) para rotular o mamífero insetívoro identificado no item anterior com o seu nome, **sem cobrir outros pontos do gráfico**. Cuidado para não alterar a legenda que já existe.

4.10.2

Peso cerebral e horas de sono

Use o *data frame* `sono` definido como

```
library(ggplot2)

sono <- msleep %>%
  select(
    name, order, genus, vore, bodywt,
    brainwt, awake, sleep_total
  )
```

1. Construa um histograma da variável `brainwt`. Escolha o número de classes que você achar melhor. O que acontece com os valores `NA`?
2. Descubra que função da forma `scale_x_... usar` para fazer com que o eixo *x* tenha uma escala logarítmica. Gere um novo histograma.
3. Qual dos dois histogramas é melhor para responder a pergunta “Qual a faixa de peso cerebral que tem mais animais?” de forma satisfatória?
4. Construa um *scatter plot* de horas de sono versus peso do cérebro. Você percebe alguma correlação entre estas variáveis? Se precisar, concentre-se em um subconjunto dos dados.
5. Usando `geom_smooth` ([leia a respeito](#)), sobreponha uma reta de regressão ao gráfico de dispersão, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.

4.10.3

Igualdade de gênero entre furacões?

[Este artigo](#) tenta achar uma relação entre o gênero do nome de um furacão e a quantidade de vítimas fatais provocadas por ele.

Os dados estão no pacote `DAAG`, que deve ser instalado:

```
if (!require(DAAG))
  install.packages("DAAG")
```

Vamos usar apenas algumas das variáveis, com nomes em português.

```
df <- hurricNamed %>%
  as_tibble() %>%
  transmute(
    id = paste(Year, Name, sep = '-'),
    nome = Name,
    ano = Year,
    velocidade = LF.WindsMPH * 1.8,      # convertido para km/h
    pressao = LF.PressureMB,             # mbar
    prejuizo = BaseDam2014 %>% round(), # milhões de dólares de 2014
    mortes = deaths,
    genero = mf
  )
```

1. Crie histogramas para as seguintes variáveis, escolhendo a quantidade de barras que você achar melhor.

- velocidade
- prejuizo
- mortes

Não se esqueça de incluir títulos nos gráficos e rótulos nos eixos.

Comente os histogramas.

2. Os histogramas de prejuízos e mortes não ficaram bons. Vamos gerar histogramas transformados.

No *data frame*, crie duas novas colunas:

- logprejuizo: *logaritmo* do prejuízo (na base 10)
- logmortes: *logaritmo* do número de mortes (na base 10)

Agora, gere histogramas destas duas novas variáveis.

3. O que significa o valor do logaritmo do prejuízo na base 10?
4. O que significa o valor do logaritmo do número de mortes na base 10?
5. Por que o histograma do logaritmo do número de mortes vem com uma mensagem de aviso?
6. Por que isto não acontece com o logaritmo do prejuízo?
7. Faça um gráfico de dispersão com *pressao* no eixo *y* e *velocidade* no eixo *x*.
8. Usando `geom_smooth` ([leia a respeito](#)), sobreponha uma reta de regressão ao gráfico, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.
9. Faça um gráfico de dispersão com *logmortes* no eixo *y* e *pressao* no eixo *x*.

10. Usando `geom_smooth` ([leia a respeito](#)), sobreponha uma reta de regressão ao gráfico, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.
11. Faça um gráfico de dispersão com `logmortes` no eixo y e `pressao` no eixo x , com pontos coloridos de acordo com o gênero do nome do furacão.
12. Usando `geom_smooth` ([leia a respeito](#)), sobreponha retas de regressão ao gráfico, uma para cada gênero, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.



Visualizações como esta ajudam a explorar os dados, mas não servem para testar rigorosamente a hipótese de que furacões matam mais do que furacões homens.

Mais adiante no curso, vamos aprender a fazer testes mais rigorosos sobre hipóteses como esta.

Visualização com ggplot2 (continuação)



Busque mais informações sobre os pacotes `tidyverse` e `ggplot2` nas referências recomendadas.

5.1

Vídeo 1

<https://youtu.be/TjgLDeIQHIc>

5.2

Boxplots

5.2.1

Conjunto de dados

- Vamos continuar a trabalhar com os dados sobre as horas de sono de alguns mamíferos:

```
sono <- msleep %>%  
  select(name, vore, order, sleep_total)
```

```
sono
```

```
## # A tibble: 83 x 4
##   name                vore order      sleep_total
##   <chr>              <chr> <chr>         <dbl>
## 1 Cheetah            carni Carnivora      12.1
## 2 Owl monkey         omni  Primates        17
## 3 Mountain beaver    herbi Rodentia       14.4
## 4 Greater short-tailed shrew omni  Soricomorpha  14.9
## 5 Cow                herbi Artiodactyla    4
## 6 Three-toed sloth    herbi Pilosa      14.4
## # ... with 77 more rows
```

5.2.2

Mediana e quartis

- Para entender *boxplots*, precisamos, antes, entender algumas medidas.
- Se tomarmos as quantidades de horas de sono de todos os animais do conjunto de dados e **classificarmos estas quantidades em ordem crescente**, vamos ter:

```
horas <- sono %>%
  pull(sleep_total) %>%
  sort()
```

```
horas
```

```
## [1] 1,9 2,7 2,9 3,0 3,1 3,3 3,5 3,8 3,9 4,0 4,4 5,2 5,3
## [14] 5,3 5,4 5,6 6,2 6,3 6,3 7,0 7,7 8,0 8,3 8,4 8,4 8,6
## [27] 8,7 8,7 8,9 9,1 9,1 9,4 9,4 9,5 9,6 9,7 9,8 9,8 10,0
## [40] 10,1 10,1 10,1 10,3 10,3 10,4 10,6 10,9 11,0 11,0 11,1 11,3 11,5
## [53] 12,1 12,5 12,5 12,5 12,5 12,8 12,8 13,0 13,5 13,7 13,8 14,2 14,3
## [66] 14,4 14,4 14,5 14,6 14,9 14,9 15,6 15,8 15,8 15,9 16,6 17,0 17,4
## [79] 18,0 18,1 19,4 19,7 19,9
```

- Quantos valores são?

```
length(horas)
```

```
## [1] 83
```

- O valor que está **bem no meio desta fila** — i.e., na posição 42 — é a **mediana**:

```
horas[ceiling(length(horas) / 2)]
```

```
## [1] 10,1
```

- Em R:

```
median(horas)
```

```
## [1] 10,1
```

Mediana e média são coisas muito diferentes.

Por acaso, neste exemplo, a média das horas é próxima da mediana:



```
mean(horas)
```

```
## [1] 10,43373
```

Isto costuma acontecer quando a distribuição dos dados é aproximadamente simétrica.

- Os **quartis** são os valores que estão nas posições $\frac{1}{4}$, $\frac{1}{2}$ e $\frac{3}{4}$ da fila. São o **primeiro**, **segundo e terceiro quartis**, respectivamente.

```
horas[  
  c(  
    ceiling(length(horas) / 4),  
    ceiling(length(horas) / 2),  
    ceiling(3 * length(horas) / 4)  
  )  
]
```

```
## [1] 7,7 10,1 13,8
```

- **Sim, a mediana é o segundo quartil.**
- Em R, a **função quantile** generaliza esta ideia: dado um número q entre 0 e 1, o **quantil (com “N”) q é o elemento que está na posição que corresponde à fração q da fila ordenada.**

```
horas %>% quantile(c(.25, .5, .75))
```

```
## 25% 50% 75%
```

```
## 7,85 10,10 13,75
```

- Na verdade, R tem 9 algoritmos diferentes para calcular os quantis de uma amostra! Leia a ajuda da função `quantile` para conhecê-los.
- As diferenças entre nossos cálculos “à mão” e os resultados retornados por `quantile` são porque, em algumas situações, `quantile` calcula uma média ponderada entre elementos vizinhos. Por isso, `quantile` pode retornar valores que nem estão no vetor.
- Em R, a **função summary** mostra o **mínimo**, os **quartis (com “R”)**, a **média**, e o **máximo** de um vetor:


```
summary(horas)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1,90    7,85   10,10   10,43   13,75   19,90
```

5.2.3

Média × mediana

- Vamos ver um exemplo simples para entender a diferença entre a média e a mediana.
- Imagine o seguinte vetor com as receitas mensais de algumas pessoas (em milhares de reais:)

```
receitas <- c(1, 2, 2, 3.5, 1, 4, 1)
```

- Eis a mediana e a média deste vetor:

```
summary(receitas)[c('Median', 'Mean')]
```

```
##      Median      Mean
## 2,000000 2,071429
```

- A mediana e a média são bem próximas.
- Imagine, agora, que adicionamos ao vetor um sujeito com receita mensal de 100 mil reais:

```
receitas <- c(1, 2, 2, 3.5, 1, 4, 1, 100)
```

- Eis a nova mediana e a nova média:

```
summary(receitas)[c('Median', 'Mean')]
```

```
##      Median      Mean
## 2,0000 14,3125
```

- O sujeito com a receita de 2 mil reais continua no meio da fila, mas a média (que é a soma de todas as receitas, dividida pelo número de indivíduos) ficou muito diferente.
- A receita do novo sujeito é um **valor discrepante**, ou, em inglês, um **outlier**.

Conclusão:



A **mediana é robusta**, pouco afetada por *outliers*.

A **média é pouco robusta**, muito sensível a *outliers*.

5.2.4

Intervalo interquartil (IQR) e *outliers*

- Qual fração dos elementos está entre o primeiro e o terceiro quartis?

```
length(
  horas[between(horas, quantile(horas, .25), quantile(horas, .75))]
) /
length(
  horas
)
```

```
## [1] 0,4939759
```

- Metade do total de elementos está entre o primeiro e o terceiro quartis.
- Este é o chamado intervalo interquartil (*interquartile range*, em inglês).
- No nosso vetor `horas`, os limites do IQR são

```
quantile(horas, c(.25, .75))
```

```
## 25% 75%
```

```
## 7,85 13,75
```

- O comprimento deste intervalo é calculado pela função `IQR`:

```
IQR(horas)
```

```
## [1] 5,9
```

- Valores muito abaixo do primeiro quartil podem ser considerados discrepantes (*outliers*), mas quão abaixo?
- A resposta (puramente convencional) é $1,5 \times \text{IQR}$ abaixo do primeiro quartil.
- No nosso vetor `horas`, isto significa valores abaixo de

```
limite_inferior <- quantile(horas, .25) - 1.5 * IQR(horas)

unnname(limite_inferior)
```

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

- Neste caso, não há *outliers*:

```
horas[horas < limite_inferior]
```

```
## numeric(0)
```

- Da mesma forma, valores **muito acima do terceiro quartil** podem ser considerados discrepantes (*outliers*), mas quão acima?
- De novo, a resposta (puramente convencional) é **$1,5 \times \text{IQR}$ acima do terceiro quartil**.
- No nosso vetor `horas`, isto significa valores acima de

```
limite_superior <- quantile(horas, .75) + 1.5 * IQR(horas)

unnamed[limite_superior]
```

```
## [1] 22,6
```

- Neste caso, também não há *outliers*:

```
horas[horas > limite_superior]
```

```
## numeric(0)
```

- Outro exemplo: vamos tomar apenas os mamíferos onívoros:

```
onivoros <- sono %>%
  filter(vore == 'omni')

onivoros
```

```
## # A tibble: 20 x 4
##   name                vore order      sleep_total
##   <chr>              <chr> <chr>         <dbl>
## 1 Owl monkey        omni  Primates      17
## 2 Greater short-tailed shrew omni  Soricomorpha 14.9
## 3 Grivet            omni  Primates      10
## 4 Star-nosed mole    omni  Soricomorpha 10.3
## 5 African giant pouched rat omni  Rodentia       8.3
## 6 Lesser short-tailed shrew omni  Soricomorpha  9.1
## # ... with 14 more rows
```

- Vamos extrair o vetor de horas de sono:

```
horas <- onivoros %>%
  pull(sleep_total)

horas
```

```
## [1] 17,0 14,9 10,0 10,3  8,3  9,1 18,0 10,1 10,9  9,8  8,0 10,1  9,7
## [14]  9,4 11,0  8,7  9,6  9,1 15,6  8,9
```

- Vamos calcular o primeiro e terceiro quartis:

```
quartis <- horas %>%
  quantile(c(.25, .75))

quartis
```

```
##      25%      75%
##  9,100 10,925
```

- Vamos achar o IQR:

```
IQR(horas)
```

```
## [1] 1,825
```

- E os limites a partir dos quais os valores são *outliers*:

```
limites <- quartis + c(-1, 1) * 1.5 * IQR(horas)

unnname(limites)
```

```
## [1] 6,3625 13,6625
```

- Existem *outliers* inferiores?

```
onivoros %>%
  filter(sleep_total < limites[1])
```

```
## # A tibble: 0 x 4
## #   ... with 4 variables: name <chr>, vore <chr>, order <chr>,
## #   sleep_total <dbl>
```

Não.

- Existem *outliers* superiores?

```
onivoros %>%
  filter(sleep_total > limites[2])
```

```
## # A tibble: 4 x 4
##   name                vore order      sleep_total
##   <chr>              <chr> <chr>          <dbl>
## 1 Owl monkey        omni  Primates         17
## 2 Greater short-tailed shrew omni  Soricomorpha    14.9
## 3 North American Opossum  omni  Didelphimorphia  18
## 4 Tenrec             omni  Afrosoricida    15.6
```

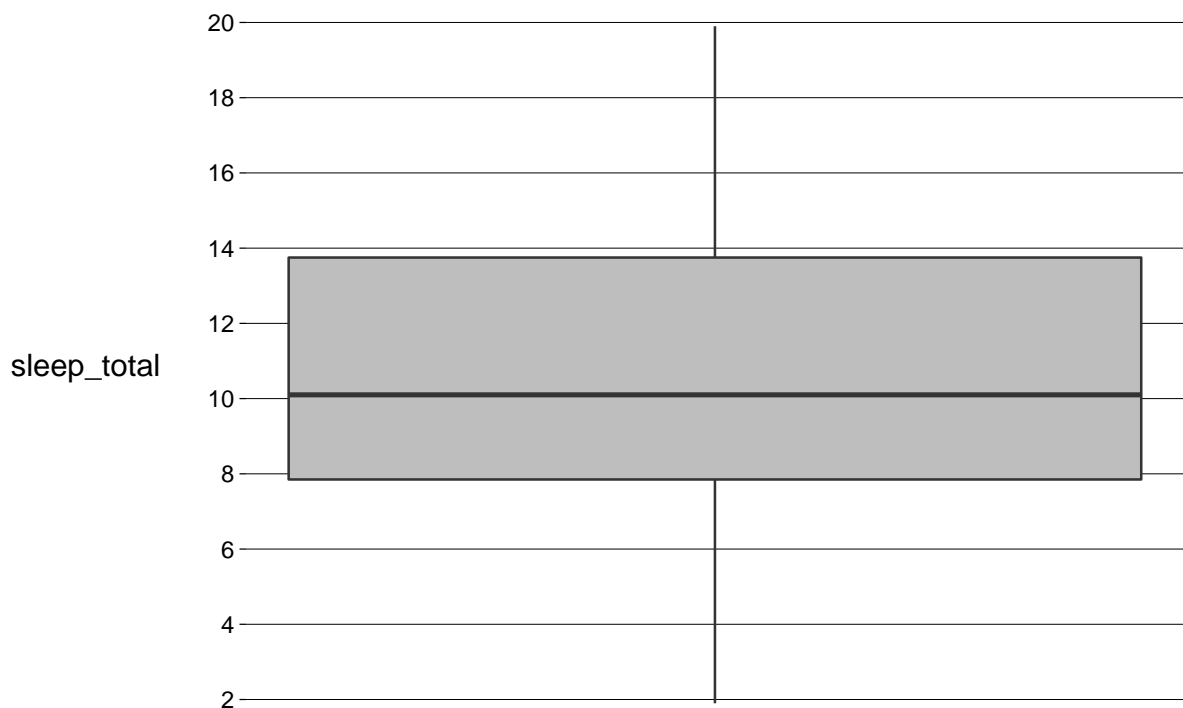
Sim! Estes animais dormem demais em comparação com os outros onívoros.

5.2.5

Gerando boxplots

- Um *boxplot* é uma representação visual dos valores que calculamos acima.
- No `ggplot2`, a geometria `geom_boxplot` constrói *boxplots*:

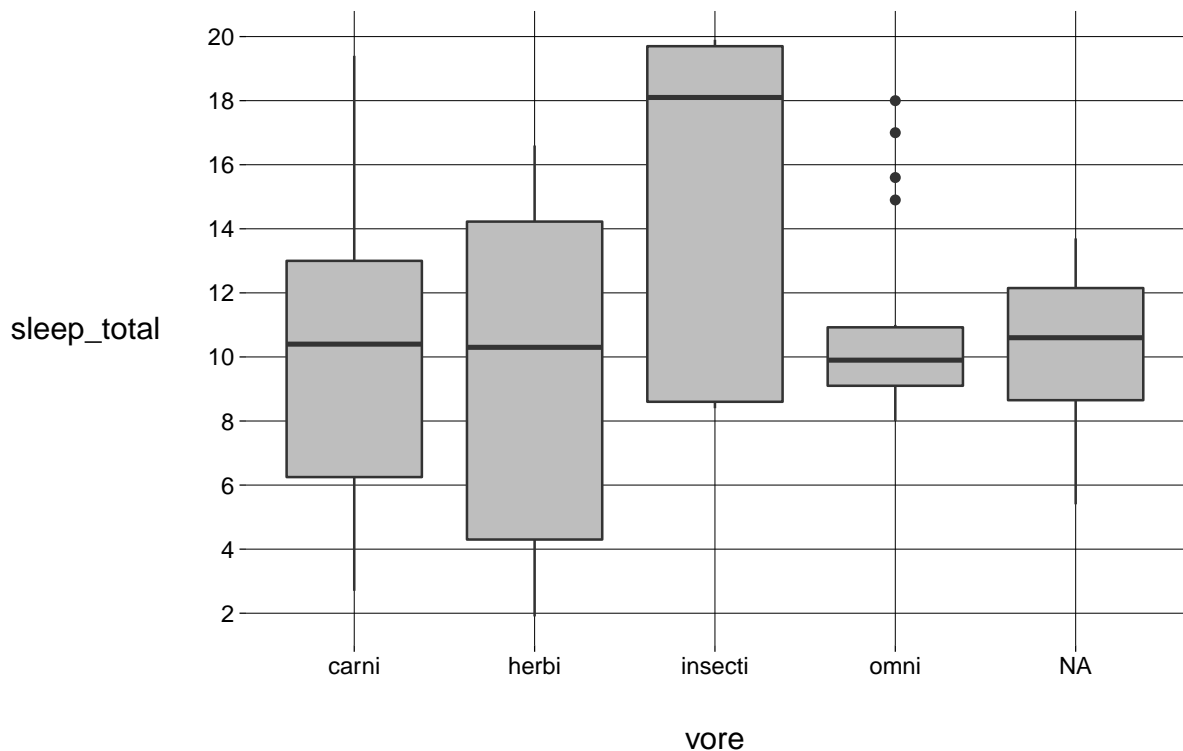
```
sono %>%  
  ggplot(aes(y = sleep_total)) +  
    geom_boxplot(fill = 'gray') +  
    scale_x_continuous(breaks = NULL) +  
    scale_y_continuous(breaks = seq(0, 20, 2))
```



- A **caixa** vai do valor do **primeiro quartil** (embaixo) até o **terceiro quartil** (em cima).
- A **linha horizontal dentro da caixa** representa o valor da **mediana**.
- As **linhas verticais** acima e abaixo da caixa (pitorescamente chamadas de “bigodes”) vão até o **limite inferior** (primeiro quartil $- 1,5 \times \text{IQR}$) e até o **limite superior** (terceiro quartil $+ 1,5 \times \text{IQR}$).
- Neste *boxplot*, não há *outliers*.
- Podemos usar a posição *x* para desenhar vários *boxplots*, um para cada dieta:

```
sono %>%  
  ggplot(aes(x = vore, y = sleep_total)) +
```

```
geom_boxplot(fill = 'gray') +
scale_y_continuous(breaks = seq(0, 20, 2))
```



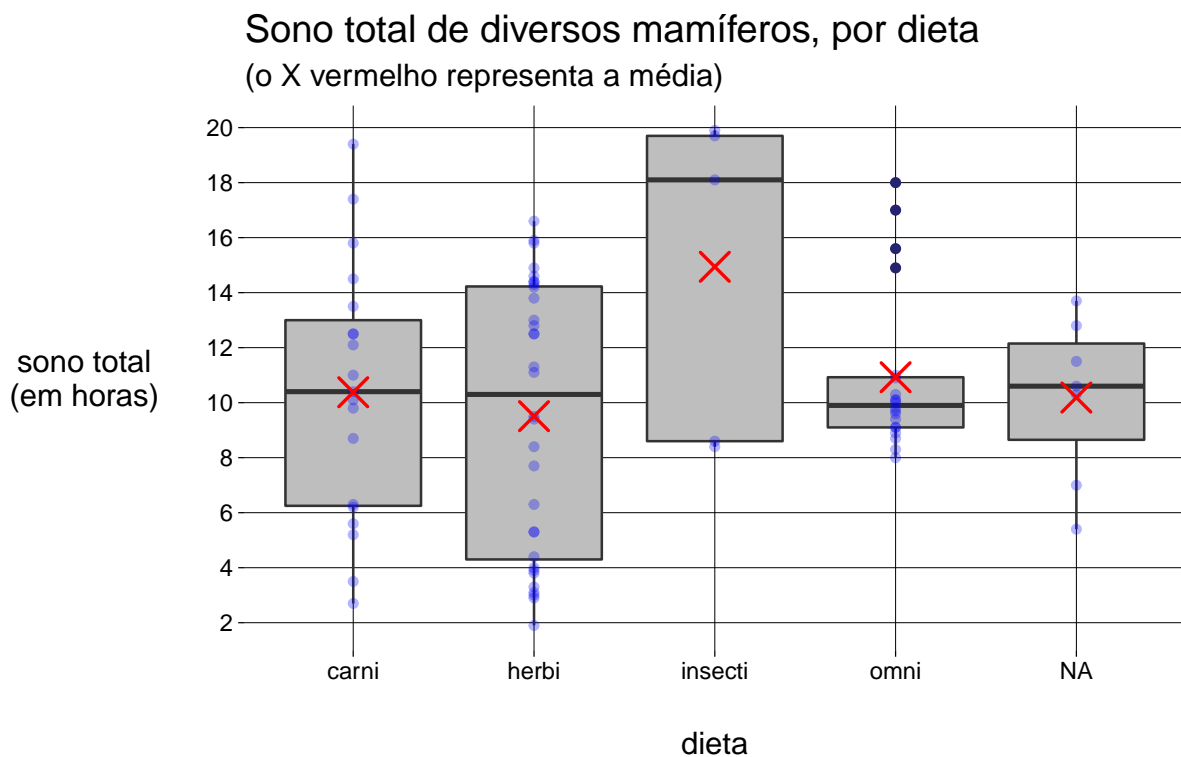
- No *boxplot* de onívoros, **os outliers aparecem como pontos isolados**, acima da caixa, além dos alcances do bigode superior (aliás, onde está bigode superior?).
- *Boxplots* lado a lado são úteis para compararmos grupos diferentes de dados.
- Veja como, com exceção dos insetívoros, as medianas dos grupos são parecidas.
- Veja como carnívoros, insetívoros e herbívoros apresentam maior variação, enquanto onívoros e animais sem dieta registrada apresentam menor variação.
- Vamos combinar, em um só gráfico
 - Os pontos representando os animais,
 - Os *boxplots*,
 - As médias (que podem estar próximas ou distantes das medianas).

```
sono %>%
  ggplot(aes(x = vore, y = sleep_total)) +
    geom_boxplot(fill = 'gray') +
    scale_y_continuous(breaks = seq(0, 20, 2)) +
    geom_point(
      color = 'blue',
      alpha = .3
    ) +
```

```

stat_summary(
  fun = mean,
  geom = 'point',
  color = 'red',
  shape = 'cross',
  size = 5,
  stroke = 1
) +
labs(
  title = 'Sono total de diversos mamíferos, por dieta',
  subtitle = '(o X vermelho representa a média)',
  x = 'dieta',
  y = 'sono total\n(em horas)'
)

```



- Quando a caixa é longa, o IQR é grande, e os valores estão muito espalhados; é o caso dos herbívoros e insetívoros.
- Quando a caixa é curta, o IQR é pequeno, e os valores estão pouco espalhados; é o caso dos onívoros. Como o IQR é pequeno, os 4 mamíferos com mais de 14 horas de sono são *outliers*.
- Observe, ainda, como os *outliers* “puxam” a média dos onívoros para cima.

5.3

Vídeo 2

<https://youtu.be/QqnOvgBXJ-s>

5.4

Gráficos de barras e de colunas

5.4.1

Conjunto de dados

- O R tem um *array* de 3 dimensões com dados sobre as cores dos cabelos e dos olhos de 592 alunos e alunas de uma universidade americana em 1974.
- Se pedirmos para o R exibir os dados, veremos **duas matrizes**, uma para cada sexo:

```
HairEyeColor
```

```
## , , Sex = Male
##
##      Eye
## Hair   Brown Blue Hazel Green
## Black   32   11   10    3
## Brown   53   50   25   15
## Red     10   10    7    7
## Blond    3   30    5    8
##
## , , Sex = Female
##
##      Eye
## Hair   Brown Blue Hazel Green
## Black   36    9    5    2
## Brown   66   34   29   14
## Red     16    7    7    7
## Blond    4   64    5    8
```

- Vamos transformar este *array* em um *data frame*.
- O *array* contém apenas os totais de cada classe. Vamos usar a função `uncount` para gerar uma linha para cada aluno:

```
df_orig <- as.data.frame(HairEyeColor) %>%
  uncount(Freq) %>%
  as_tibble()
```



```
df_orig
```

```
## # A tibble: 592 x 3
##   Hair Eye   Sex
##   <fct> <fct> <fct>
## 1 Black Brown Male
## 2 Black Brown Male
## 3 Black Brown Male
## 4 Black Brown Male
## 5 Black Brown Male
## 6 Black Brown Male
## # ... with 586 more rows
```

- O `ggplot2` e os outros pacotes do `tidyverse` foram projetados para trabalhar com *data frames* neste formato, com uma observação (um indivíduo, um elemento) por linha. É o chamado *formato tidy*.
- Usando vetores com elementos nomeados, podemos traduzir o conteúdo do *data frame* para português:

```
cabelo <- c(
  'Brown' = 'castanhos',
  'Blond' = 'loiros',
  'Black' = 'pretos',
  'Red' = 'ruivos'
)

olhos <- c(
  'Brown' = 'castanhos',
  'Blue' = 'azuis',
  'Hazel' = 'avelã',
  'Green' = 'verdes'
)

sexo <- c(
  'Male' = 'homem',
  'Female' = 'mulher'
)

df <- df_orig %>%
  transmute(
    cabelos = cabelo[Hair],
    olhos = olhos[Eye],
    sexo = sexo[Sex]
  )
```

- Um sumário:

```
df %>% dfSummary() %>% print()
```

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
cabelos	1. castanhos	108 (18,2%)	0
[character]	2. louros	286 (48,3%)	(0,0%)
	3. pretos	71 (12,0%)	
	4. ruivos	127 (21,5%)	
olhos	1. avelã	93 (15,7%)	0
[character]	2. azuis	215 (36,3%)	(0,0%)
	3. castanhos	220 (37,2%)	
	4. verdes	64 (10,8%)	
sexo	1. homem	279 (47,1%)	0
[character]	2. mulher	313 (52,9%)	(0,0%)

5.4.2

Gerando gráficos de barras

- Um **gráfico de barras** contém uma barra para cada valor de uma **variável categórica**.
- Usamos **geom_bar** para gerar um gráfico de barras de cores de cabelo:

```
df %>%
  ggplot(aes(x = cabelos)) +
  geom_bar() +
  labs(y = NULL)
```

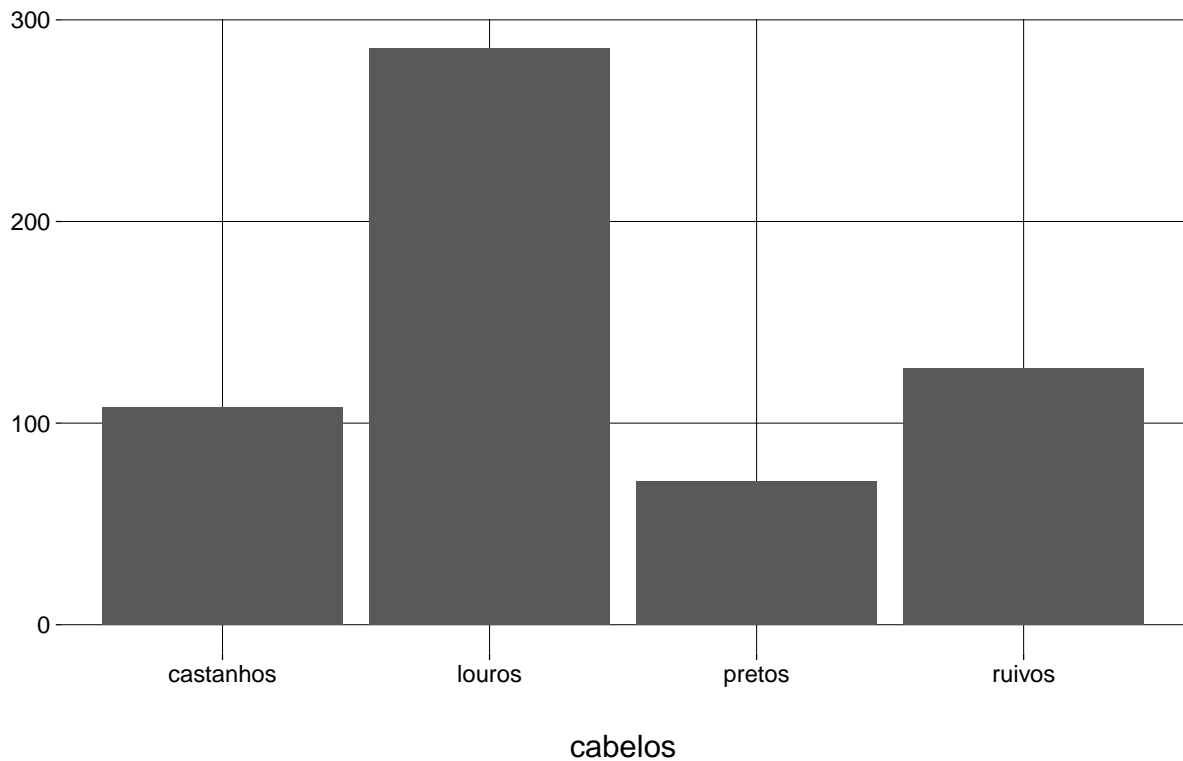


Gráfico de barras × histograma:

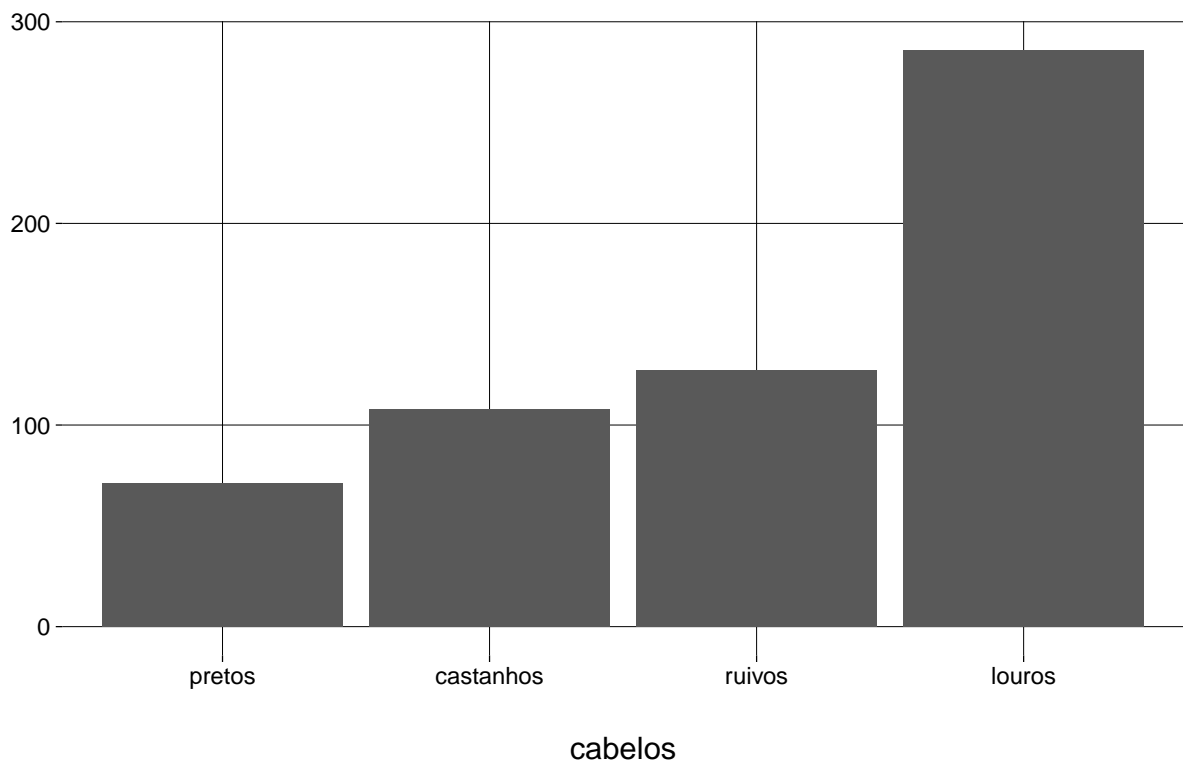
- Os dois tipos de gráficos mostram a frequência (quantidade de elementos) no eixo vertical.
- No gráfico de barras:
 - * A variável é categórica (nominal).
 - * Cada barra corresponde a um valor da variável.
 - * As barras não se tocam, enfatizando o fato de que a variável é categórica.
- No histograma (veja o exemplo):
 - * A variável é quantitativa (intervalar ou racional).
 - * Cada barra corresponde a uma classe de valores da variável.
 - * As barras se tocam, para enfatizar que as classes são contíguas.

- Um gráfico de barras é mais legível quando as barras são mostradas em ordem crescente ou decrescente.
- Embora os valores da variável `cabelos` sejam *strings*, podemos aplicar a eles funções que manipulam fatores.
- A função `fct_infreq`, do pacote `forcats`, ordena os valores em ordem decrescente

de frequência.

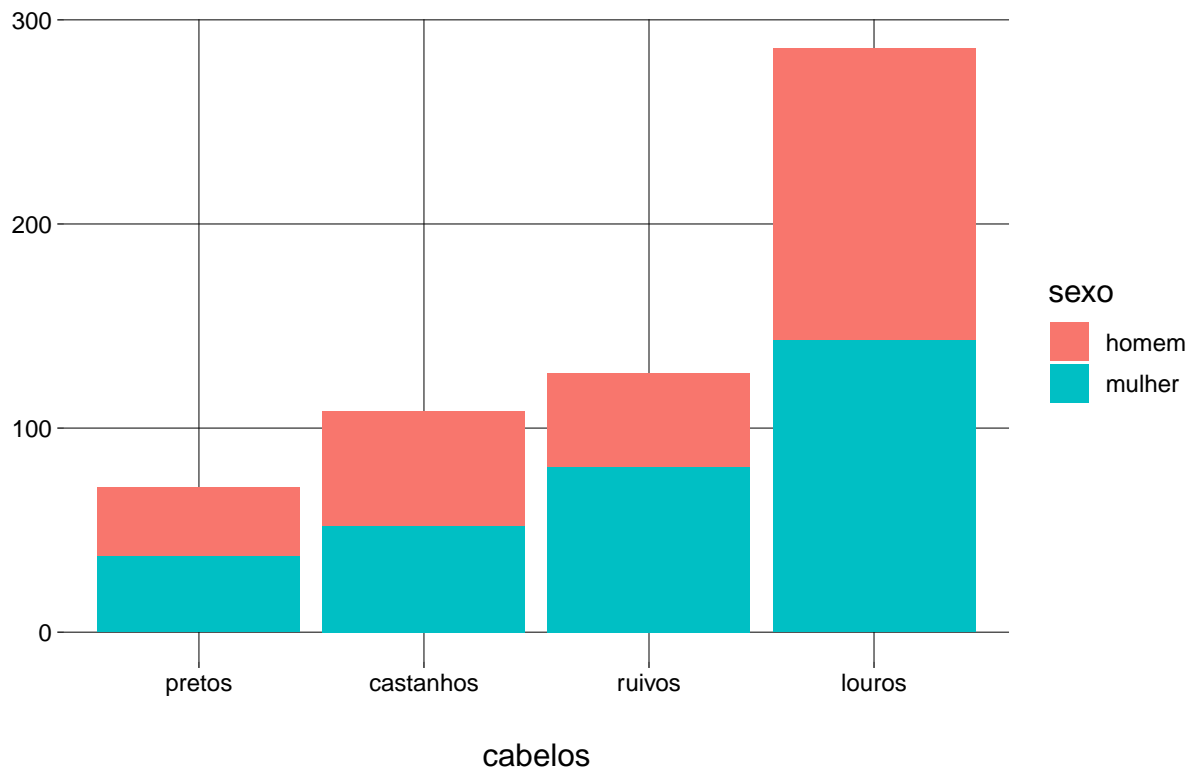
- A função `fct_rev`, também do pacote `forcats`, **inverte a ordenação**.

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)))) +  
    geom_bar() +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    )
```



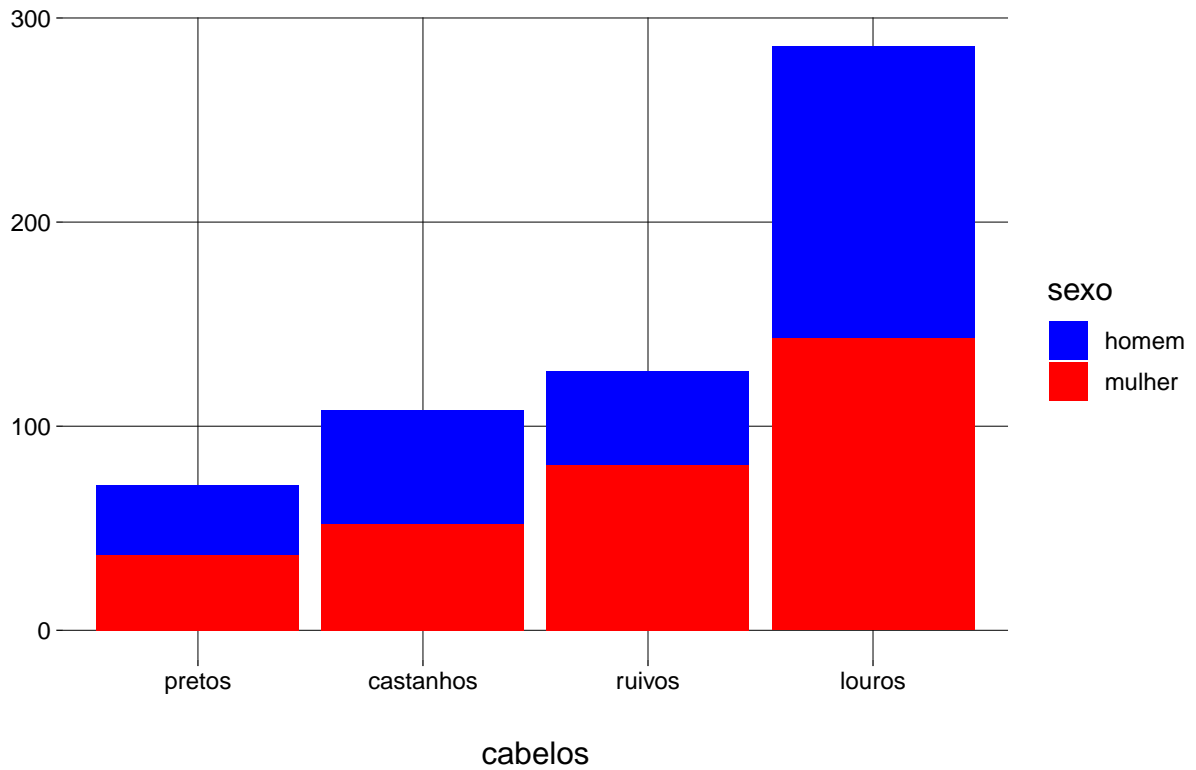
- A posição x e a altura de cada barra são estéticas: **a posição x representa a cor dos cabelos**, e **a altura representa a frequência daquela cor**.
- Vamos acrescentar mais uma estética: **a cor de preenchimento vai representar o sexo**.

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +  
    geom_bar() +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    )
```



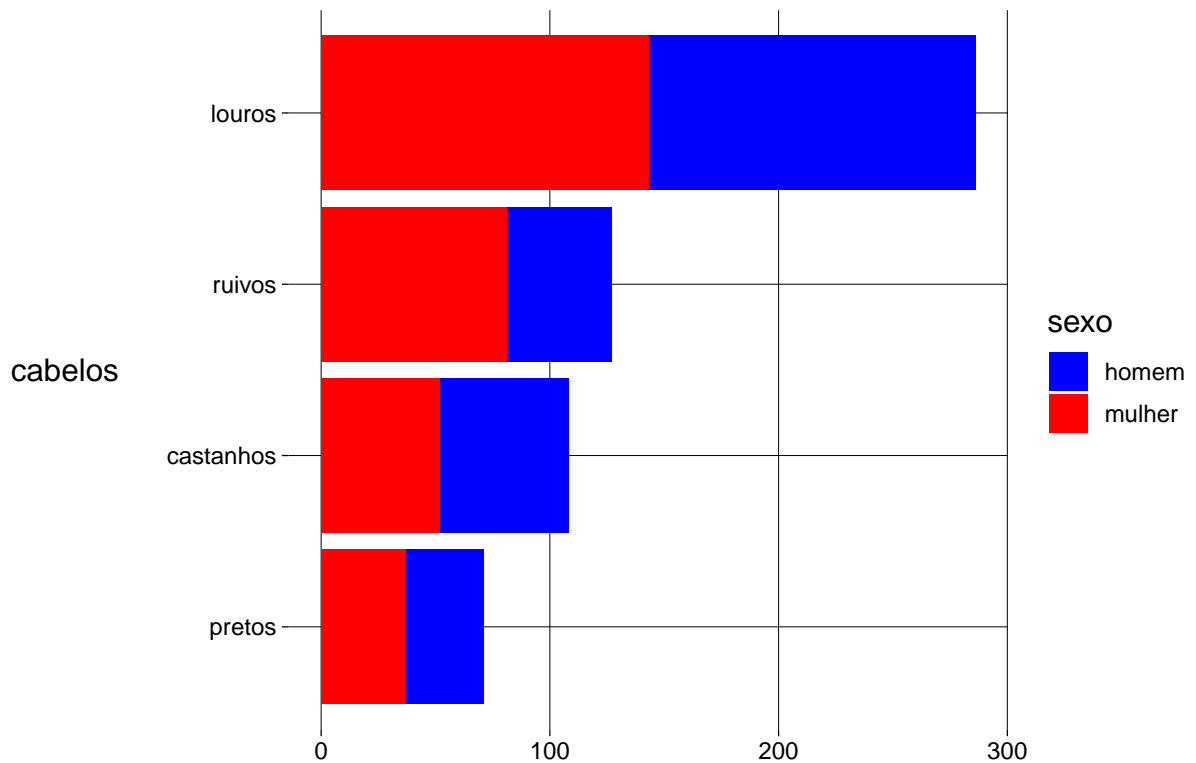
- Se a cor dos homens incomoda você, altere a escala que especifica o preenchimento (scale_fill_discrete):

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +  
    geom_bar() +  
    scale_fill_discrete(type = c('blue', 'red')) +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    )
```



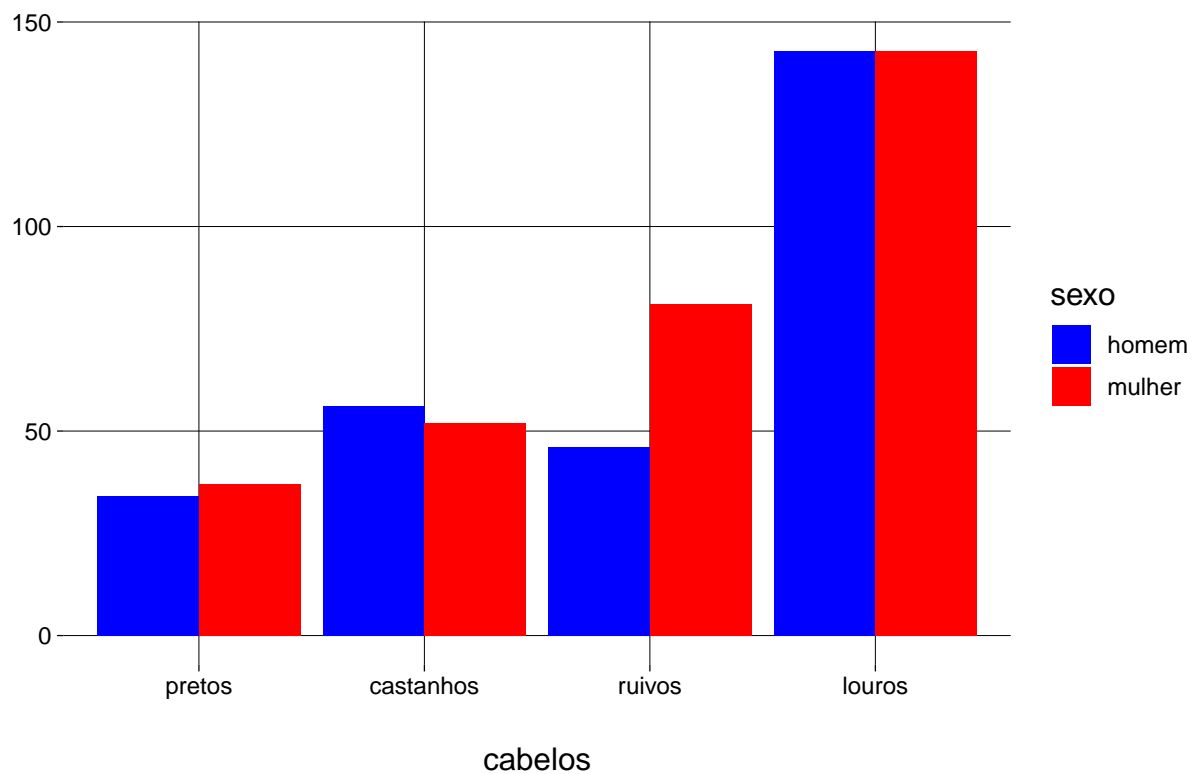
- Podemos fazer um gráfico de barras horizontais com `coord_flip`. Isto geralmente é útil quando os rótulos das barras são longos:

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +  
    geom_bar() +  
    scale_fill_discrete(type = c('blue', 'red')) +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    ) +  
    coord_flip()
```



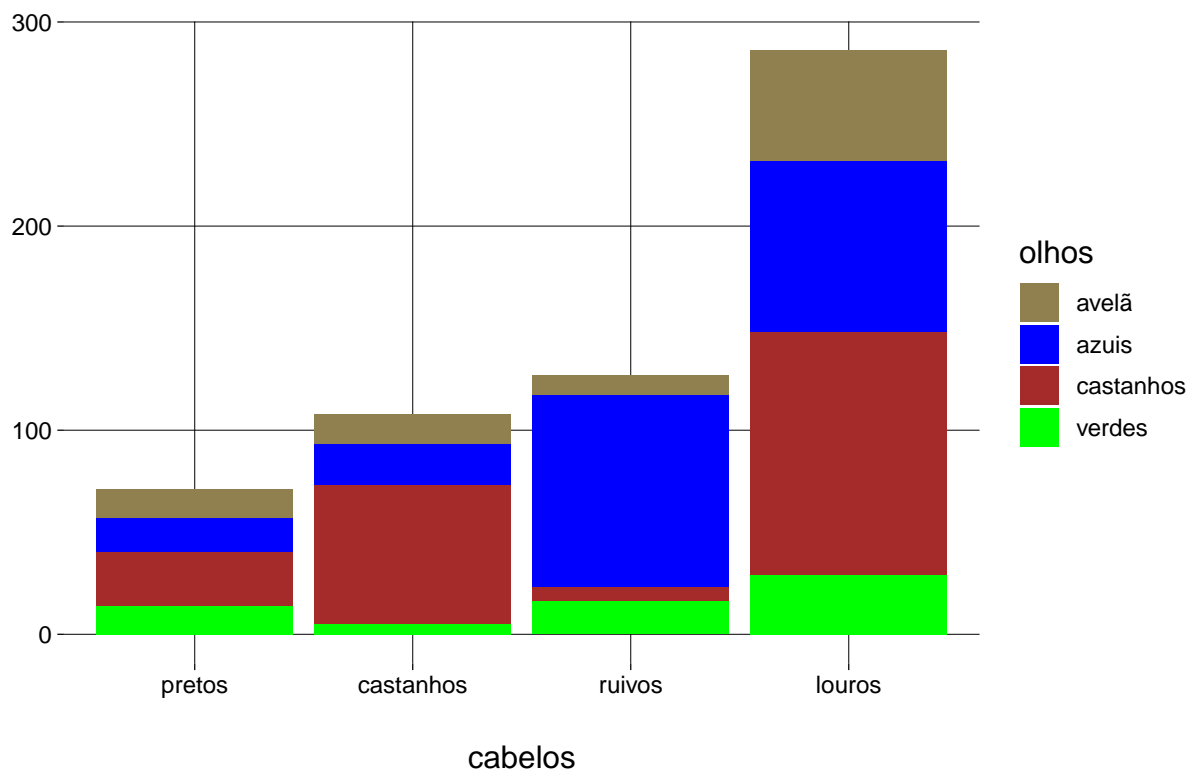
- Você consegue dizer se há mais homens ou mulheres com cabelos pretos? E castanhos? E ruivos?
- Se posicionarmos as barras lado a lado, fica mais fácil responder.
- Usamos o argumento `position = 'dodge'` de `geom_bar`. “*Dodge*” significa “esquivar-se”, em inglês.

```
df %>%
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +
  geom_bar(position = 'dodge') +
  labs(
    x = 'cabelos',
    y = NULL
  ) +
  scale_fill_discrete(type = c('blue', 'red'))
```



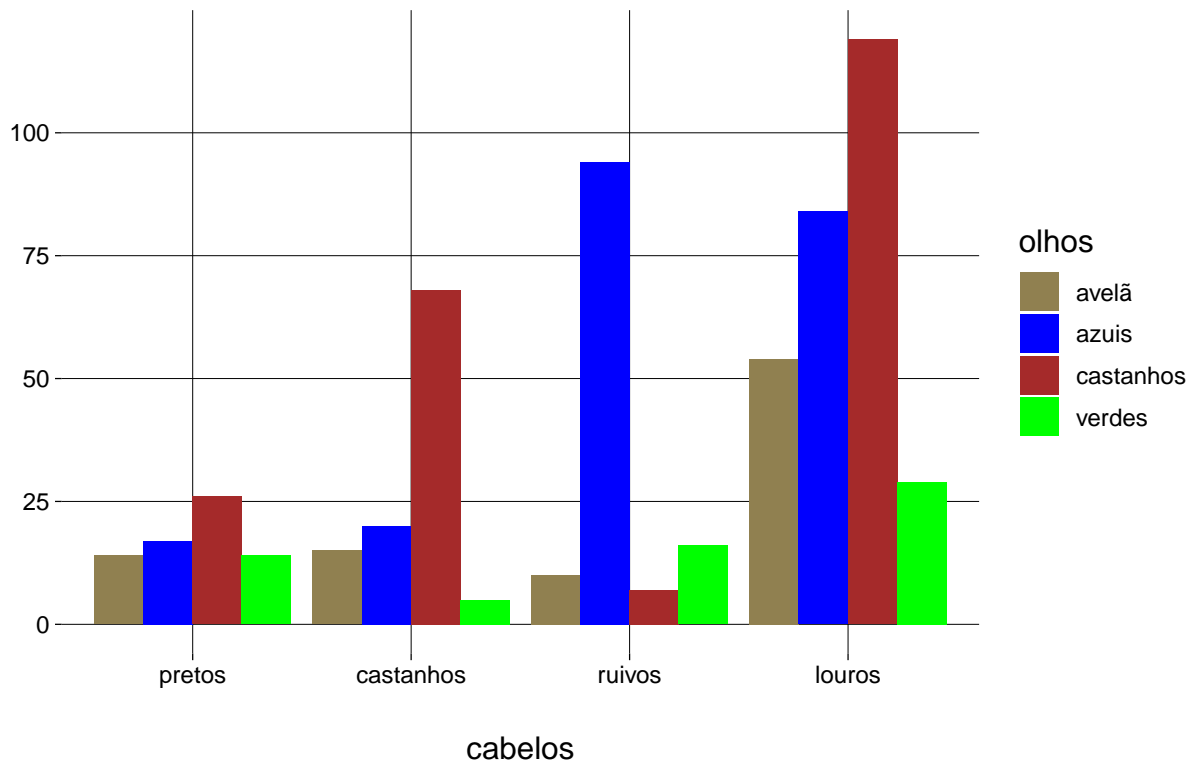
- Agora vamos examinar a relação entre as cores dos olhos e as cores dos cabelos:

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = olhos)) +  
    geom_bar() +  
    scale_fill_discrete(  
      type = c('#908050', 'blue', 'brown', 'green')  
    ) +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    )
```

- Ou, com barras lado a lado:

```
df %>%
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = olhos)) +
  geom_bar(position = 'dodge') +
  scale_fill_discrete(
    type = c('#908050', 'blue', 'brown', 'green')
  ) +
  labs(
    x = 'cabelos',
    y = NULL
  )
```



- Observações e perguntas:

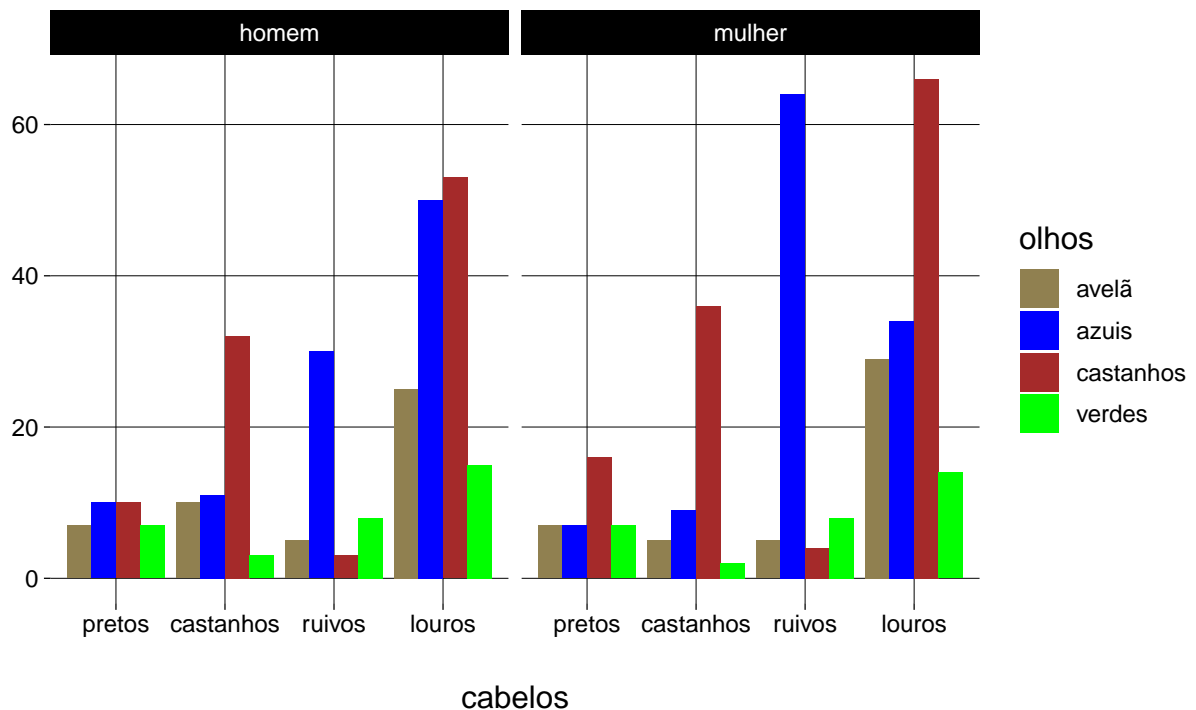
1. Há mais pessoas loiras de olhos castanhos do que loiras de olhos azuis. O esperado não seria mais pessoas loiras de olhos azuis? Pessoas loiras de olhos castanhos pintaram os cabelos?
2. Há muito mais ruivos de olhos azuis do que ruivos de olhos verdes. Não deveria ser o contrário? Também são pessoas que pintaram os cabelos de ruivo? Ou houve erro no registro das cores dos olhos?

- Para incluir o sexo, podemos **facetar** o gráfico. Usando `facet_wrap`¹, geramos dois subgráficos lado a lado:

```
df %>%
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = olhos)) +
  geom_bar(position = 'dodge') +
  scale_fill_discrete(type = c('#908050', 'blue', 'brown', 'green')) +
  facet_wrap(~sexo) +
  labs(
    title = 'Cores de cabelos e olhos por sexo',
    y = NULL,
    x = 'cabelos'
  )
```

¹O nome da variável segundo a qual facetar deve aparecer depois de um ~.

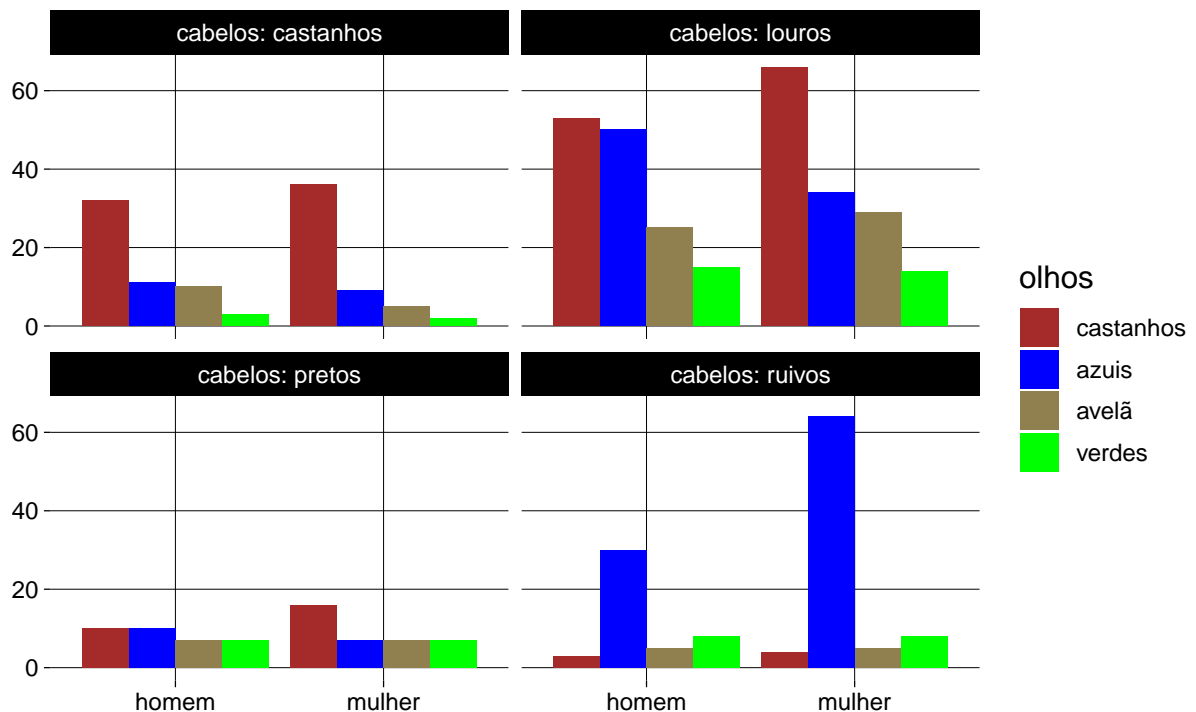
Cores de cabelos e olhos por sexo



- Se a quantidade grande de pessoas louras de olhos castanhos (em comparação com pessoas louras de olhos azuis) for por causa da pintura de cabelos, então o gráfico acima mostra que as mulheres pintam os cabelos de louro com mais frequência do que os homens.
- Quando facetamos por cor de cabelos, também podemos observar as mesmas diferenças entre homens e mulheres:

```
df %>%
  ggplot(aes(x = sexo, fill = fct_infreq(olhos))) +
  geom_bar(position = 'dodge') +
  facet_wrap(~cabelos, labeller = label_both) +
  scale_fill_discrete(type = c('brown', 'blue', '#908050', 'green')) +
  labs(
    x = NULL,
    y = NULL,
    fill = 'olhos',
    title = 'Cor dos olhos e sexo por cor dos cabelos'
  )
```

Cor dos olhos e sexo por cor dos cabelos



5.4.3

Data frame já contendo os totais

- Você percebeu que `geom_bar` analisa o *data frame* e calcula as frequências necessárias para construir o gráfico.
- Em algumas situações, o *data frame* já contém as frequências (em vez de conter uma linha por indivíduo).
- Vamos usar `count` para criar um *data frame* assim:

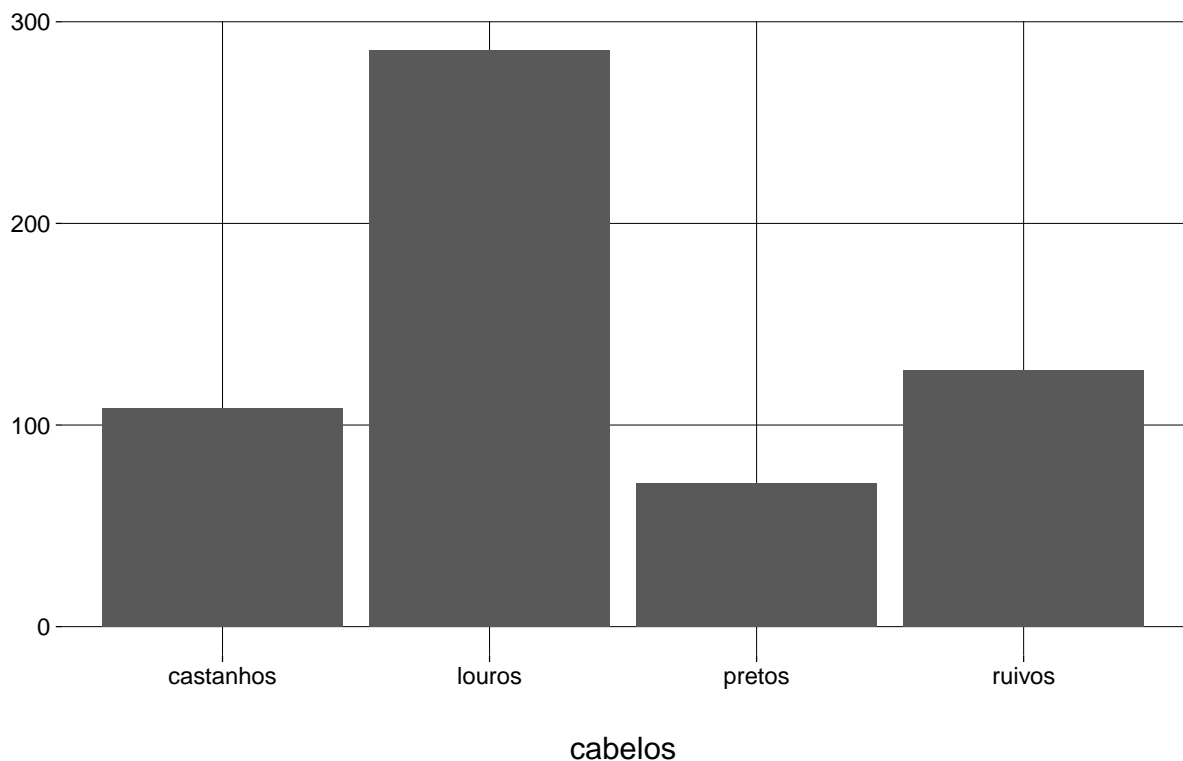
```
df_tot <- df %>%
  count(sexo, cabelos, olhos)
```

```
df_tot
```

```
## # A tibble: 32 x 4
##   sexo  cabelos  olhos      n
##   <chr> <chr>    <chr>  <int>
## 1 homem castanhos avelã    10
## 2 homem castanhos azuis    11
## 3 homem castanhos castanhos  32
## 4 homem castanhos verdes     3
## 5 homem louros    avelã    25
## 6 homem louros    azuis    50
## # ... with 26 more rows
```

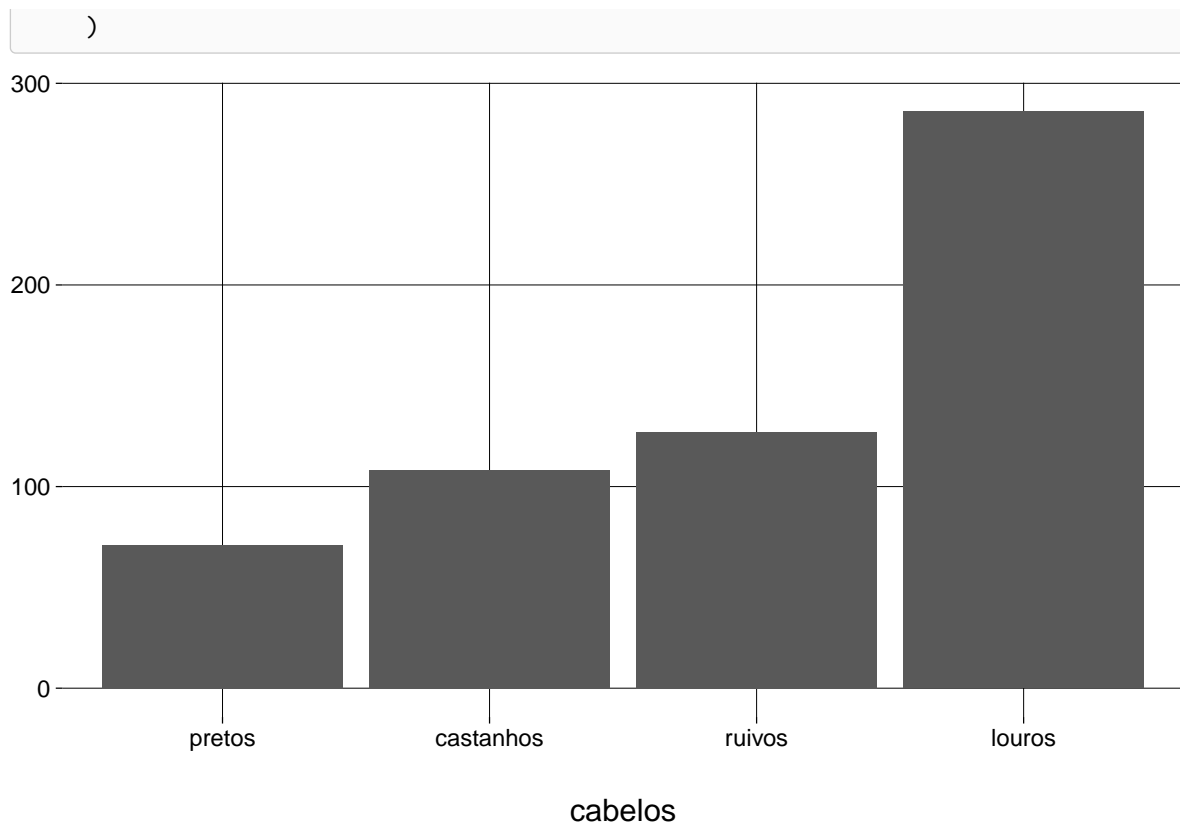
- Para 4 cores de cabelo, 4 cores de olhos, e 2 sexos, são 32 combinações possíveis.
- Com este *data frame*, podemos gerar todos os gráficos anteriores usando `geom_col` no lugar de `geom_bar`. Por exemplo:

```
df_tot %>%
  ggplot(aes(x = cabelos, y = n)) +
    geom_col() +
    labs(
      y = NULL
    )
```



- Com `geom_col`, precisamos passar a estética `y` (no nosso exemplo, a variável `n`, que contém as frequências).
- Para ordenar as barras, usamos a função `fct_reorder`, que ordena os níveis de um fator (`cabelos`) de acordo com o resultado de uma função (`sum`) aplicada sobre os valores de outra variável (`n`):

```
df_tot %>%
  ggplot(aes(x = fct_reorder(cabelos, n, sum), y = n)) +
    geom_col() +
    labs(
      x = 'cabelos',
      y = NULL
    )
```



5.5

Gráficos de linha e séries temporais

5.5.1

Conjunto de dados

- O R tem uma matriz com as quantidades de telefones em várias regiões do mundo ao longo de vários anos:

WorldPhones

##		N.Amer	Europe	Asia	S.Amer	Oceania	Africa	Mid.Amer
##	1951	45939	21574	2876	1815	1646	89	555
##	1956	60423	29990	4708	2568	2366	1411	733
##	1957	64721	32510	5230	2695	2526	1546	773
##	1958	68484	35218	6662	2845	2691	1663	836
##	1959	71799	37598	6856	3000	2868	1769	911
##	1960	76036	40341	8220	3145	3054	1905	1008
##	1961	79831	43173	9053	3338	3224	2005	1076

- Os números representam milhares.
- Os números dos anos são os nomes das linhas da matriz.

- Vamos transformar esta matriz em uma *tibble*:

```
fones <- WorldPhones %>%
  as_tibble(rownames = 'Ano') %>%
  mutate(Ano = as.numeric(Ano))
```

```
fones
```

```
## # A tibble: 7 x 8
##   Ano N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1951 45939 21574 2876 1815 1646 89 555
## 2 1956 60423 29990 4708 2568 2366 1411 733
## 3 1957 64721 32510 5230 2695 2526 1546 773
## 4 1958 68484 35218 6662 2845 2691 1663 836
## 5 1959 71799 37598 6856 3000 2868 1769 911
## 6 1960 76036 40341 8220 3145 3054 1905 1008
## # ... with 1 more row
```

- Esta *tibble* não está no formato *tidy*. Queremos que cada linha corresponda a uma observação, contendo
 - Ano,
 - Região,
 - Quantidade de telefones.
- Usamos a função `pivot_longer` para mudar o formato da *tibble*:

```
fones_long <- fones %>%
  pivot_longer(
    cols = -Ano,
    names_to = 'Região',
    values_to = 'n'
  )
```

```
fones_long
```

```
## # A tibble: 49 x 3
##   Ano Região      n
##   <dbl> <chr> <dbl>
## 1 1951 N.Amer 45939
## 2 1951 Europe 21574
## 3 1951 Asia 2876
## 4 1951 S.Amer 1815
## 5 1951 Oceania 1646
## 6 1951 Africa 89
## # ... with 43 more rows
```

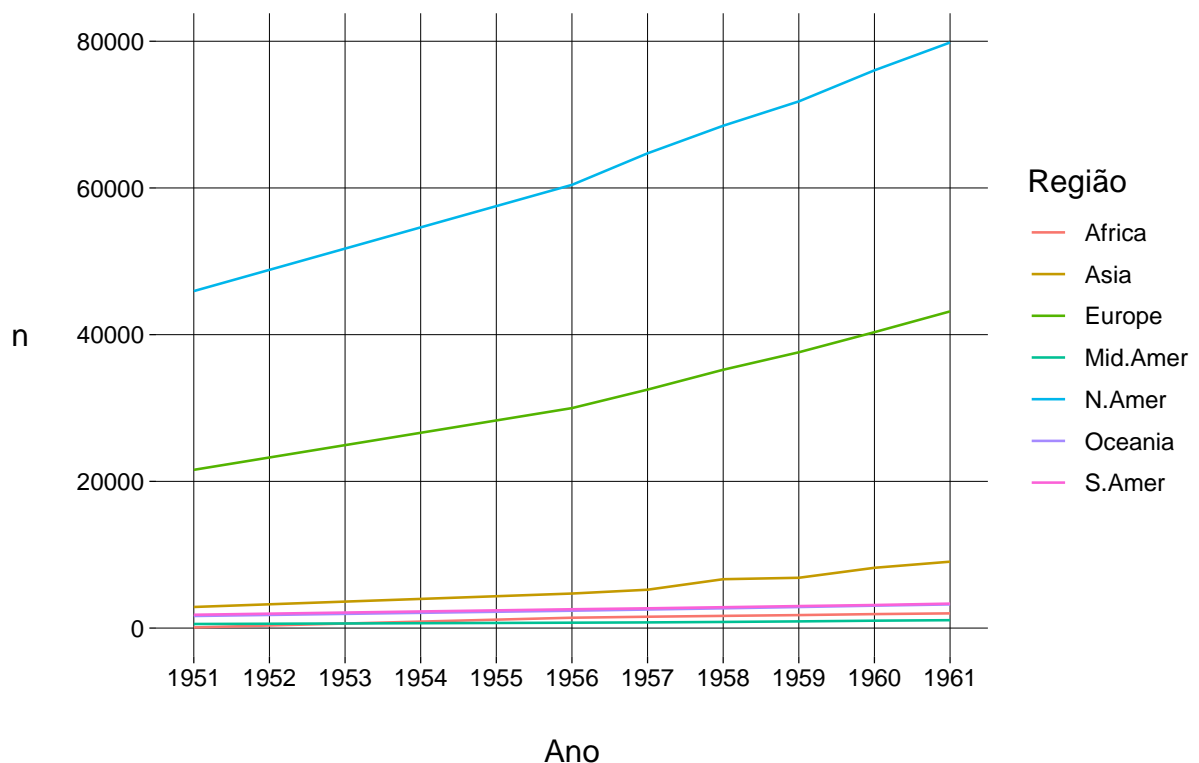
- Confira: antes, tínhamos 7 anos, com 7 quantidades por ano, uma quantidade por região. Eram 49 quantidades. Agora temos uma *tibble* de 49 linhas.

5.5.2

Gerando gráficos de linha

- A geometria `geom_line` gera gráficos de linha. Perceba como geramos uma linha por região:

```
fones_long %>%
  ggplot(aes(x = Ano, y = n, color = Região)) +
    geom_line() +
    scale_x_continuous(breaks = 1951:1961)
```



- Embora a legenda associe uma cor a cada região, a leitura seria mais fácil se a ordem das regiões na legenda coincidissem com a posição das linhas na borda direita da grade:

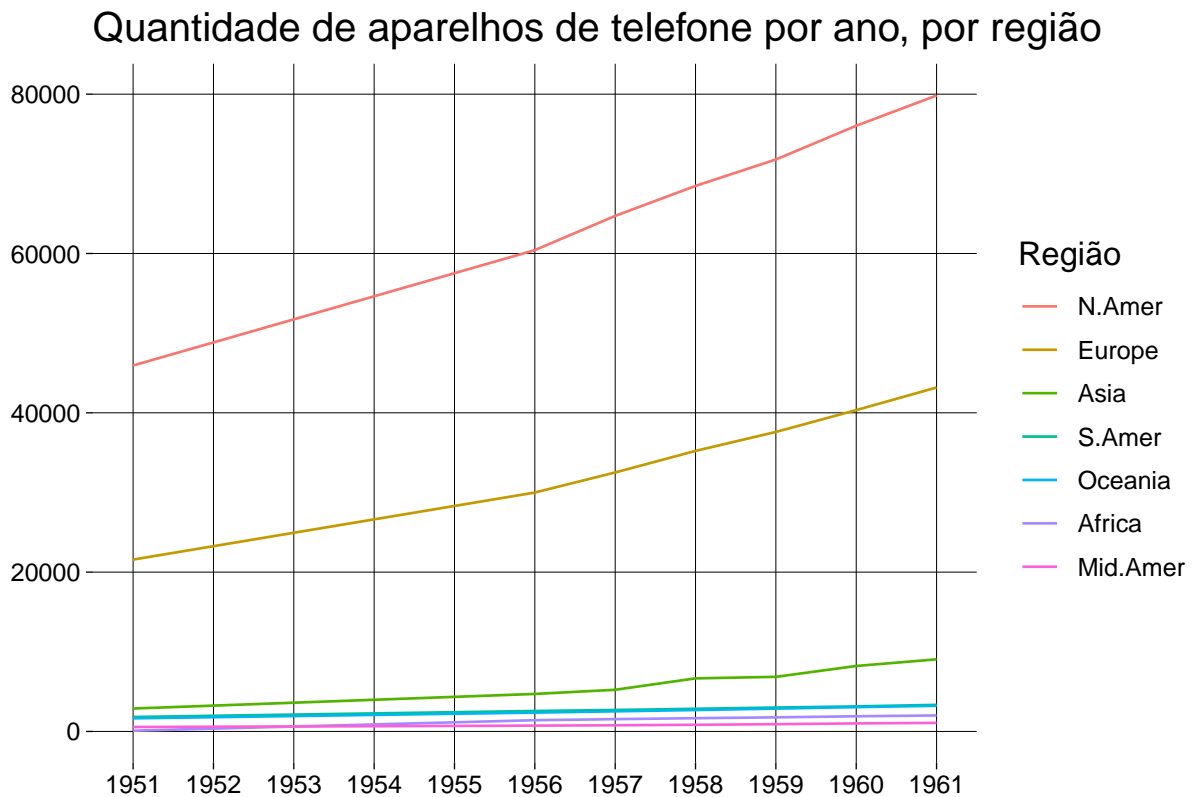
```
fones_long %>%
  ggplot(
    aes(
      x = Ano,
      y = n,
```



```

    color = fct_rev(fct_reorder(Região, n, max))
  )
) +
  geom_line() +
  scale_x_continuous(breaks = 1951:1961) +
  labs(
    color = 'Região',
    y = '',
    x = NULL,
    title = 'Quantidade de aparelhos de telefone por ano, por região'
  )

```



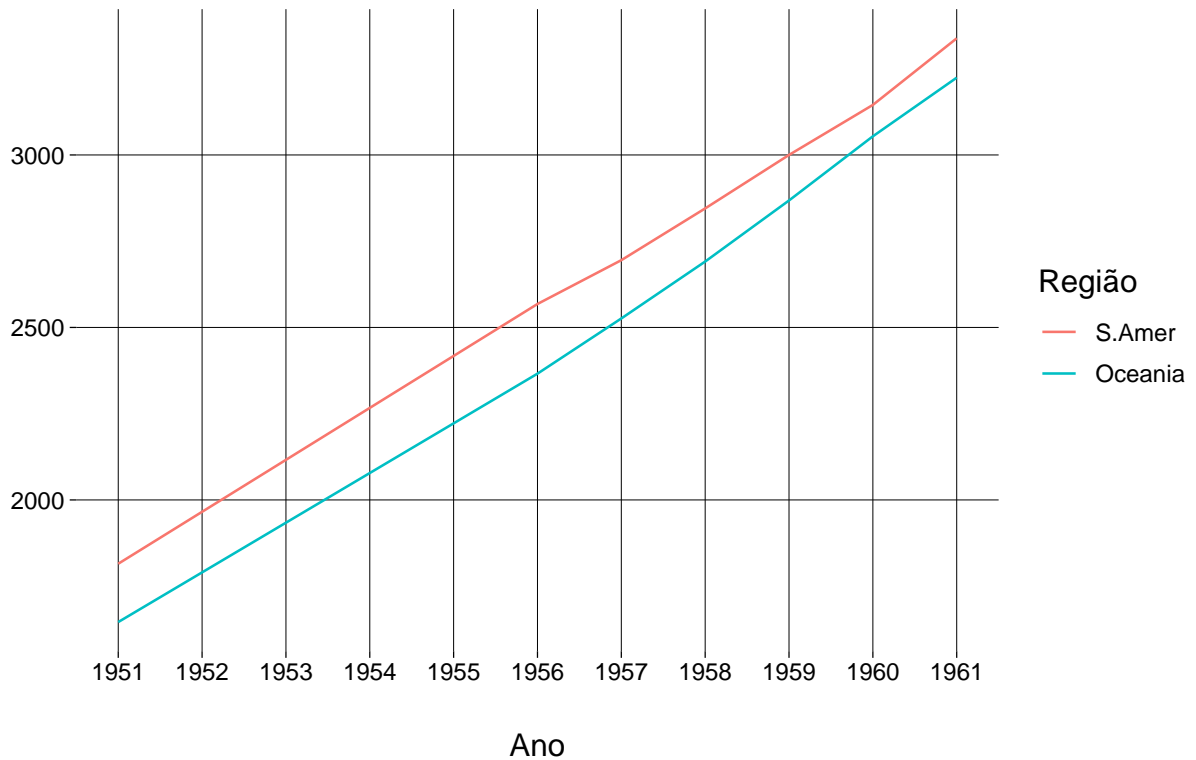
- Parece que está faltando uma linha, mas o que acontece é que as quantidades da América do Sul e da Oceania são bem parecidas:

```

fones_long %>%
  filter(Região %in% c('S.Amer', 'Oceania')) %>%
  ggplot(
    aes(
      x = Ano,
      y = n,
      color = fct_rev(fct_reorder(Região, n, max))
    )
  ) +

```

```
geom_line() +
  scale_x_continuous(breaks = 1951:1961) +
  labs(y = NULL, color = 'Região')
```



- Estamos tratando estes dados como simples números, mas, na verdade, **este conjunto de dados é uma série temporal (*time series*)**.
- R tem todo um conjunto de funções para tratar séries temporais, calcular tendências, achar padrões cíclicos, fazer estimativas, e gerar gráficos específicos, entre outras coisas.
- Mas não vamos falar mais sobre séries temporais aqui.
- O **pacote tsibble** oferece maneiras de trabalhar com séries temporais de maneira *tidy*. Você pode ler a documentação do pacote entrando

```
library(tsibble)
?tsibble-package
```

5.6

Exercícios

5.6.1

O bigode dos onívoros

- Examine o *data frame* `sono` para descobrir o que houve com o bigode superior do *boxplot* dos onívoros [neste gráfico](#).

5.6.2

Usando `geom_col`

- Use `geom_col` para reproduzir, a partir do *data frame* `df_tot`, todos os gráficos que foram gerados com `geom_bar` na seção [Gerando gráficos de barras](#).

5.7

Referências sobre visualização e R



Busque mais informações sobre os pacotes `tidyverse` e `ggplot2` [nas referências recomendadas](#).

Medidas

6.1

Vídeo

<https://youtu.be/C96MOP4YlaY>

6.2

Medidas de centralidade

6.2.1

Média

- A **média de uma população** é escrita como μ , e é definida como

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

- $\sum_{i=1}^N x_i$ é a soma de todos os dados da população.
- N é a quantidade de elementos na população.

- A **média de uma amostra** é escrita como \bar{x} , e é definida como:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- $\sum_{i=1}^n x_i$ é a soma de todos os dados da amostra.
- n é a quantidade de elementos na amostra.
- O cálculo é essencialmente o mesmo. Só mudam os símbolos: N versus n , e μ versus \bar{x} .

6.2.1.1

Exemplo

- Idades dos alunos de uma turma:

```
idades <- c(
  20, 20, 20, 20, 20, 20, 21, 21, 21, 21,
  22, 22, 22, 23, 23, 23, 23, 24, 24,
  65
)
```

- Média **com** o velhinho de 65 anos:

```
mean(idades)
```

```
## [1] 23,75
```

- Média **sem** o velhinho:

```
mean(idades[-length(idades)])
```

```
## [1] 21,57895
```

6.2.2

Mediana

- Já aprendemos sobre a mediana na [seção sobre boxplots](#).
- A ideia é que, depois de ordenar os dados, 50% dos dados estarão à esquerda da mediana, e 50% à direita.
- A mediana não é tão sensível a *outliers* quanto à média.

6.2.2.1

Exemplo

- Mediana **com** o velhinho:

```
median(idades)
```

```
## [1] 21,5
```

- Mediana **sem** o velhinho:

```
median(idades[-length(idades)])
```

```
## [1] 21
```

6.2.3

Moda

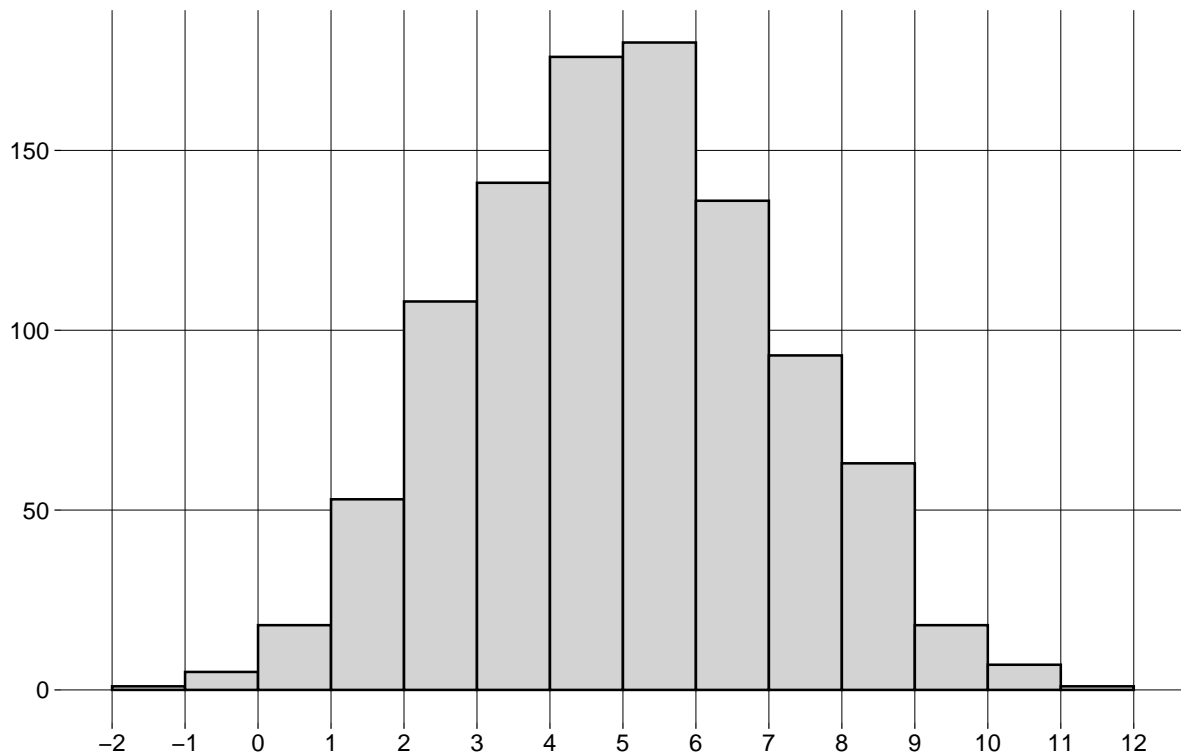
- A **moda** é o **valor mais frequente** do conjunto de dados.
- Pode haver mais de uma moda.
- **Não existe uma função para a moda em R base.** Por quê?
- Por incrível que pareça, **é complicado definir a moda de forma a conseguir resultados interessantes.**
- **Por exemplo**, vamos definir um conjunto de 1000 valores numéricos distribuídos normalmente¹, com média igual a 5 e desvio-padrão² igual a 2:

```
normal <- rnorm(1000, mean = 5, sd = 2)
```

- O histograma dos nossos dados é

¹Mais sobre a distribuição normal no capítulo ???.

²Mais sobre o desvio-padrão daqui a pouco.



- Vamos calcular a moda com a função `mfv` (*most frequent value*), do pacote `modeest`:

```
# Pacote com funções para calcular modas
library(modeest)
```

```
## Registered S3 method overwritten by 'rmutil':
##   method      from
##   print.response httr
```

```
# Por causa de um bug na função mfv,
# precisamos de números com ponto decimal
# (em vez de vírgula):
options(OutDec = '.')
mfv(normal)
```

```
##      [1] -1.66945492 -0.27624682 -0.13713112 -0.12094119 -0.06536332
##      [6] -0.01795423  0.13609223  0.33289959  0.34804058  0.35762175
##     [11]  0.38340677  0.43580706  0.53682352  0.69176975  0.72966165
##     [16]  0.76217920  0.82304019  0.83023861  0.88669790  0.90302511
##     [21]  0.90657207  0.90803862  0.91771026  0.99430211  1.00161332
##     [26]  1.00405813  1.04463933  1.05889354  1.06600620  1.09597331
##     [31]  1.09737843  1.10894352  1.13956902  1.14285443  1.16611248
##     [36]  1.17570281  1.17610675  1.17942339  1.20165899  1.20882109
##     [41]  1.21077652  1.21699705  1.23633069  1.25155765  1.26430127
##     [46]  1.28336633  1.30382838  1.31968343  1.32178380  1.36873775
##     [51]  1.41459736  1.43683951  1.43859834  1.45754064  1.46691225
```

##	[56]	1.49482076	1.51097287	1.53275076	1.53520959	1.66285296
##	[61]	1.66858985	1.68644013	1.69542849	1.71313864	1.71420571
##	[66]	1.72917834	1.76151240	1.79322806	1.83714171	1.85377245
##	[71]	1.85533522	1.86680576	1.87652194	1.88560214	1.89411346
##	[76]	1.92796284	1.96810623	2.01515144	2.03103492	2.05605305
##	[81]	2.08805090	2.09629858	2.10022393	2.12184906	2.12817031
##	[86]	2.14022406	2.16789674	2.17200763	2.17860907	2.19369680
##	[91]	2.24007891	2.26761384	2.27263642	2.28552132	2.29460525
##	[96]	2.29633089	2.30773961	2.31375275	2.31695682	2.32427517
##	[101]	2.32546215	2.33242139	2.33677826	2.34760895	2.35624915
##	[106]	2.35696253	2.35743194	2.35762260	2.36442877	2.37103003
##	[111]	2.37130228	2.37591693	2.37741089	2.40775518	2.42763000
##	[116]	2.43648407	2.44926275	2.46329634	2.46400456	2.47667182
##	[121]	2.47854600	2.48895850	2.51209616	2.52047081	2.53003948
##	[126]	2.53370723	2.54467028	2.54763389	2.55503538	2.57778580
##	[131]	2.57978353	2.58622930	2.58794006	2.59570712	2.60659009
##	[136]	2.63455414	2.63458362	2.64758219	2.65300762	2.65505120
##	[141]	2.66330146	2.69027278	2.70132126	2.70372755	2.70404168
##	[146]	2.71684851	2.73069755	2.73408326	2.74063720	2.74280864
##	[151]	2.74281530	2.75160144	2.75463690	2.76382270	2.76468358
##	[156]	2.78870424	2.79471592	2.79721563	2.79782176	2.79925216
##	[161]	2.81681647	2.82865417	2.84865446	2.86267346	2.86455826
##	[166]	2.86751929	2.86972129	2.87677939	2.89180848	2.89346528
##	[171]	2.89364114	2.89888458	2.90896833	2.91277888	2.92150527
##	[176]	2.93092578	2.93753995	2.93886348	2.94661350	2.94903705
##	[181]	2.95162577	2.97392889	2.98236747	2.98590733	2.98984349
##	[186]	3.00253351	3.00763112	3.01596412	3.01825068	3.02659726
##	[191]	3.03097429	3.04901042	3.05855511	3.06081707	3.06680557
##	[196]	3.07568370	3.09838425	3.10295538	3.10495535	3.10641267
##	[201]	3.11640749	3.12351487	3.14268826	3.14914125	3.15997979
##	[206]	3.16000262	3.16852820	3.19237211	3.19515576	3.20757878
##	[211]	3.22268857	3.24027741	3.24481513	3.24491614	3.24754616
##	[216]	3.24836865	3.25083134	3.26407286	3.26653748	3.27486371
##	[221]	3.27896083	3.28654884	3.31274627	3.33127504	3.33748765
##	[226]	3.34459781	3.35594084	3.35744188	3.37429867	3.37904541
##	[231]	3.39545647	3.39780864	3.39922593	3.40885072	3.41646328
##	[236]	3.42322994	3.42518171	3.43045561	3.44433288	3.44683679
##	[241]	3.45176483	3.45493510	3.45572534	3.45692773	3.46872509
##	[246]	3.46915603	3.47014507	3.47630185	3.48389794	3.48578133
##	[251]	3.51107411	3.52197011	3.52466372	3.52827104	3.53653089
##	[256]	3.54262115	3.54506378	3.54775906	3.54885824	3.54975516
##	[261]	3.55328433	3.55378546	3.55555435	3.55791528	3.56418755
##	[266]	3.56438273	3.56813789	3.57141800	3.57619187	3.58397874
##	[271]	3.58543951	3.58698435	3.59406462	3.59744640	3.60436499
##	[276]	3.61435927	3.62350630	3.62353233	3.62466370	3.63803011
##	[281]	3.66674310	3.67642209	3.68736251	3.69233333	3.69653555
##	[286]	3.69694585	3.70433119	3.71066210	3.71790662	3.72479496
##	[291]	3.73528717	3.74127486	3.74591223	3.75544550	3.77521942

##	[296]	3.78030481	3.80242643	3.80471220	3.81146455	3.82209528
##	[301]	3.82525251	3.82767416	3.82862239	3.83593564	3.84684008
##	[306]	3.85671436	3.85821045	3.85942276	3.87017906	3.89001834
##	[311]	3.89948712	3.90234451	3.90409812	3.90560313	3.91173361
##	[316]	3.91654190	3.92099596	3.92151647	3.93155854	3.93739377
##	[321]	3.97112898	3.99295849	3.99425156	3.99548109	3.99638112
##	[326]	3.99744431	4.00705458	4.00707946	4.01619912	4.02932975
##	[331]	4.03704756	4.04238719	4.05275283	4.05332015	4.05475086
##	[336]	4.06664921	4.06830613	4.06882719	4.07217590	4.08208241
##	[341]	4.08292521	4.08510506	4.09179642	4.09659253	4.10121397
##	[346]	4.10303121	4.10622202	4.10697794	4.11257515	4.11377373
##	[351]	4.11513427	4.12654118	4.12655091	4.12906707	4.12939257
##	[356]	4.13318945	4.13408595	4.15564023	4.15634615	4.16424545
##	[361]	4.16477202	4.16823885	4.16889546	4.18244446	4.18964645
##	[366]	4.19526403	4.19655879	4.20340889	4.20760485	4.20832742
##	[371]	4.22078333	4.23438576	4.23527376	4.23565560	4.23650803
##	[376]	4.23860036	4.25259888	4.25874269	4.26317280	4.26481473
##	[381]	4.27031426	4.27512985	4.28430173	4.28514679	4.29827957
##	[386]	4.32259816	4.32358386	4.32457371	4.32658349	4.32743429
##	[391]	4.33108466	4.34207305	4.34636671	4.34662752	4.34789658
##	[396]	4.35894259	4.36102578	4.36660942	4.37532574	4.38024323
##	[401]	4.38121339	4.40102446	4.41685176	4.41979878	4.42434175
##	[406]	4.43201734	4.43799126	4.44106707	4.44115559	4.44704485
##	[411]	4.44990940	4.45762836	4.45890262	4.46934268	4.46991995
##	[416]	4.47025435	4.48122927	4.48202323	4.48817029	4.48922946
##	[421]	4.49165638	4.49769579	4.51493934	4.51854601	4.52973466
##	[426]	4.53146933	4.53149919	4.53463202	4.54695793	4.55508669
##	[431]	4.56254878	4.56663661	4.57378387	4.58451067	4.58847111
##	[436]	4.59639282	4.59980697	4.60673534	4.60905789	4.61504301
##	[441]	4.61806150	4.62459590	4.63338308	4.64758000	4.65224177
##	[446]	4.65304341	4.65600585	4.66168201	4.67500026	4.67516506
##	[451]	4.67799508	4.70265372	4.70345265	4.70685622	4.70847941
##	[456]	4.71388761	4.74119181	4.76850784	4.77051895	4.77115820
##	[461]	4.77252680	4.77444701	4.77497422	4.79018183	4.79018968
##	[466]	4.79446611	4.80046300	4.81569087	4.81575594	4.81894041
##	[471]	4.83107004	4.83123832	4.83393613	4.83998903	4.84585071
##	[476]	4.84903850	4.86073833	4.87502101	4.87638435	4.87756786
##	[481]	4.88373626	4.88728360	4.89214204	4.89405787	4.90451339
##	[486]	4.91369752	4.92190869	4.92288290	4.92747089	4.93797791
##	[491]	4.94075215	4.94695061	4.94740847	4.95162307	4.96007464
##	[496]	4.96136657	4.96535037	4.96760974	4.97046844	4.97751333
##	[501]	4.98085175	4.99470296	5.00115886	5.00149267	5.00158259
##	[506]	5.00158745	5.00165936	5.00306014	5.02171540	5.02536442
##	[511]	5.04832040	5.05202632	5.06028539	5.06793385	5.08819224
##	[516]	5.09578425	5.09955447	5.09991833	5.10716361	5.11658958
##	[521]	5.11763364	5.11961364	5.12394951	5.12444960	5.12858509
##	[526]	5.13158843	5.13181407	5.13751886	5.14490187	5.16570482
##	[531]	5.16769796	5.16804084	5.17022603	5.17186753	5.17212414

##	[536]	5.18053031	5.18558210	5.18879944	5.19397980	5.20000613
##	[541]	5.21376261	5.21612566	5.22545078	5.24289205	5.24462016
##	[546]	5.24643504	5.24759536	5.25761906	5.25823001	5.25938088
##	[551]	5.26676728	5.27285083	5.27860151	5.28740084	5.29457068
##	[556]	5.30169650	5.30276966	5.30628510	5.30696047	5.31175614
##	[561]	5.31705387	5.32469149	5.32707804	5.33166811	5.35125799
##	[566]	5.36183080	5.36329179	5.36921892	5.37667670	5.38122241
##	[571]	5.38503573	5.38678659	5.39249869	5.39320824	5.41007412
##	[576]	5.41281029	5.41310328	5.41598644	5.42692366	5.42897476
##	[581]	5.43220950	5.43239453	5.44692485	5.45391809	5.45475109
##	[586]	5.45512988	5.45626967	5.45658473	5.46030279	5.46913038
##	[591]	5.47119287	5.47142377	5.47412714	5.47699150	5.49527271
##	[596]	5.50141110	5.50397199	5.50638828	5.50935334	5.53188068
##	[601]	5.54021819	5.54176409	5.55495248	5.55815108	5.55922243
##	[606]	5.56775455	5.57756860	5.57788689	5.59157064	5.59249723
##	[611]	5.59599038	5.60483571	5.61910532	5.61934660	5.61947330
##	[616]	5.62252555	5.62393763	5.63292726	5.64126280	5.64378246
##	[621]	5.64418935	5.64581694	5.65499761	5.65886509	5.66893334
##	[626]	5.66974101	5.67135538	5.67353822	5.67389600	5.67426053
##	[631]	5.68588850	5.70251902	5.70995780	5.71395251	5.71864692
##	[636]	5.71894525	5.72014983	5.72134629	5.72145341	5.73021664
##	[641]	5.74302578	5.75859948	5.76411045	5.76414231	5.76429590
##	[646]	5.77234813	5.78888508	5.79103745	5.79406141	5.79975372
##	[651]	5.80402019	5.80682899	5.81045489	5.81762846	5.82242177
##	[656]	5.82896929	5.83551657	5.86499088	5.86804694	5.87976624
##	[661]	5.88320002	5.88359736	5.88436989	5.88595765	5.88697255
##	[666]	5.88807660	5.89236212	5.89490245	5.90212360	5.93829110
##	[671]	5.94326931	5.94339428	5.94442595	5.94654024	5.95137614
##	[676]	5.95214215	5.95351019	5.96702451	5.97441216	5.98600034
##	[681]	5.98901421	5.99438297	6.01853587	6.02475437	6.02613381
##	[686]	6.02709677	6.04101611	6.05210390	6.05357365	6.05775125
##	[691]	6.05938023	6.07446499	6.07796684	6.10353725	6.10624567
##	[696]	6.11520366	6.14507496	6.14603896	6.16086468	6.16523102
##	[701]	6.16942369	6.16990133	6.17057291	6.17269970	6.19542087
##	[706]	6.19600045	6.19876142	6.19999074	6.21107111	6.21205764
##	[711]	6.21593566	6.22019381	6.23036567	6.23296557	6.23303009
##	[716]	6.23304412	6.23464123	6.24173440	6.25189990	6.25544611
##	[721]	6.26006287	6.27046079	6.27507763	6.27508896	6.27728154
##	[726]	6.28639248	6.29000535	6.29037107	6.29550090	6.29871551
##	[731]	6.29961356	6.30267891	6.30664794	6.31190254	6.31719008
##	[736]	6.31738287	6.31892224	6.32037380	6.32638679	6.34327533
##	[741]	6.35735117	6.36825707	6.36845827	6.37507427	6.37971906
##	[746]	6.39097354	6.39900115	6.40905668	6.43976208	6.44164654
##	[751]	6.45441652	6.46478424	6.47772568	6.47963619	6.48630168
##	[756]	6.48982677	6.49619269	6.50215812	6.50712151	6.51902721
##	[761]	6.52487546	6.52542535	6.52944051	6.53662743	6.54032607
##	[766]	6.56117494	6.56480011	6.60018168	6.60162238	6.60411550
##	[771]	6.60794989	6.61208242	6.61336902	6.61814135	6.61923166

##	[776]	6.62365906	6.62542279	6.63370491	6.63886667	6.64656592
##	[781]	6.64690635	6.66713260	6.68056656	6.68371204	6.68405422
##	[786]	6.69642485	6.70354096	6.72512723	6.72573045	6.73521057
##	[791]	6.73874521	6.74567601	6.74871768	6.75633427	6.76028201
##	[796]	6.78572952	6.78749435	6.79085872	6.79616051	6.80021936
##	[801]	6.80381693	6.81822130	6.81868087	6.82842677	6.83828800
##	[806]	6.86024898	6.86142232	6.86915483	6.93126301	6.93580768
##	[811]	6.93633675	6.95035666	6.95178986	6.95369397	6.97736580
##	[816]	6.97965483	6.98549001	6.99076515	7.00183973	7.01112369
##	[821]	7.01161015	7.01275086	7.01414578	7.03315007	7.03886322
##	[826]	7.05630445	7.06350639	7.06886493	7.07019086	7.07746506
##	[831]	7.09154382	7.10429226	7.10840184	7.11548002	7.13823801
##	[836]	7.14090392	7.14761860	7.14923196	7.15127769	7.16076917
##	[841]	7.16371713	7.16697661	7.17738151	7.17802146	7.17985254
##	[846]	7.19044098	7.19349599	7.19819585	7.20140383	7.22883873
##	[851]	7.23945945	7.23987639	7.26264701	7.26267592	7.27212317
##	[856]	7.28994888	7.29257998	7.29582151	7.31693643	7.32083444
##	[861]	7.34430756	7.35000464	7.36039215	7.39221551	7.39264281
##	[866]	7.39723487	7.41665738	7.42082472	7.44798384	7.44886177
##	[871]	7.44945414	7.44954675	7.45481561	7.46020464	7.47894851
##	[876]	7.48353843	7.49570758	7.53411347	7.53909813	7.53998488
##	[881]	7.57554997	7.58114922	7.61984971	7.62671618	7.65139795
##	[886]	7.66566312	7.67605448	7.69707164	7.70832101	7.71600679
##	[891]	7.73102368	7.74094200	7.74304478	7.76807302	7.77211724
##	[896]	7.78495446	7.79288043	7.81212290	7.83598019	7.85649102
##	[901]	7.86729254	7.87509724	7.87788007	7.88657576	7.89696957
##	[906]	7.91617830	7.92415877	7.94961316	7.97741550	7.98446785
##	[911]	7.98726037	8.03781721	8.04381911	8.05731300	8.06615939
##	[916]	8.06761469	8.07027013	8.07226758	8.07490989	8.07558420
##	[921]	8.07709706	8.08670018	8.12821410	8.13082853	8.16972252
##	[926]	8.18775359	8.20566868	8.22155544	8.26873101	8.26902593
##	[931]	8.26960015	8.27825964	8.29069156	8.31856615	8.32881527
##	[936]	8.33169092	8.33826351	8.34562290	8.34566142	8.35794825
##	[941]	8.36370822	8.36692676	8.39334465	8.44856387	8.45088446
##	[946]	8.45437278	8.46929982	8.49511606	8.49607315	8.50621234
##	[951]	8.55974440	8.57313122	8.58173993	8.61380822	8.63419940
##	[956]	8.63931014	8.66239436	8.66549423	8.68046195	8.69661393
##	[961]	8.70035349	8.70036895	8.72010199	8.72426012	8.72447737
##	[966]	8.72945997	8.76080456	8.83479364	8.91256194	8.92470806
##	[971]	8.93671363	8.95499319	8.96340921	8.97618686	9.04978189
##	[976]	9.10691878	9.12094570	9.13724931	9.20423392	9.26902495
##	[981]	9.32813065	9.32850569	9.35824524	9.42105457	9.44436441
##	[986]	9.47193904	9.53115209	9.56465285	9.56605370	9.61391067
##	[991]	9.71340391	9.97879979	10.17204450	10.22457224	10.29190296
##	[996]	10.39817765	10.74717516	10.74761631	10.78792727	11.15592122

```
# Voltamos para a vírgula como separador decimal:
options(OutDec = ',')
```

- O que houve?!
- O problema é que não há valores repetidos no conjunto de dados! Por isso, todos os 1000 valores são modais.
- Uma maneira de evitar isto é definir a moda como o **centro do intervalo mais curto que contém metade dos dados**. Usamos a função `mlv` (*most likely value*):

```
moda <- mlv(normal, method = 'venter')
moda
```

```
## [1] 4,882353
```

- Esta moda estimada pode nem estar no conjunto de dados:

```
moda %in% normal
```

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

- Mas o resultado de `mlv()` é útil, pois nos diz que, embora não haja valores repetidos, valores próximos de 5 são mais frequentes, como mostra o histograma.

6.2.3.1

Exercícios

- Arredonde os valores no vetor `normal` para 2 casas decimais e ache a(s) moda(s)
 1. com a função `mfv`, e
 2. com a função `mlv`, usando o método `venter`.

Considerando o histograma, qual das respostas você prefere? Por quê?

6.3

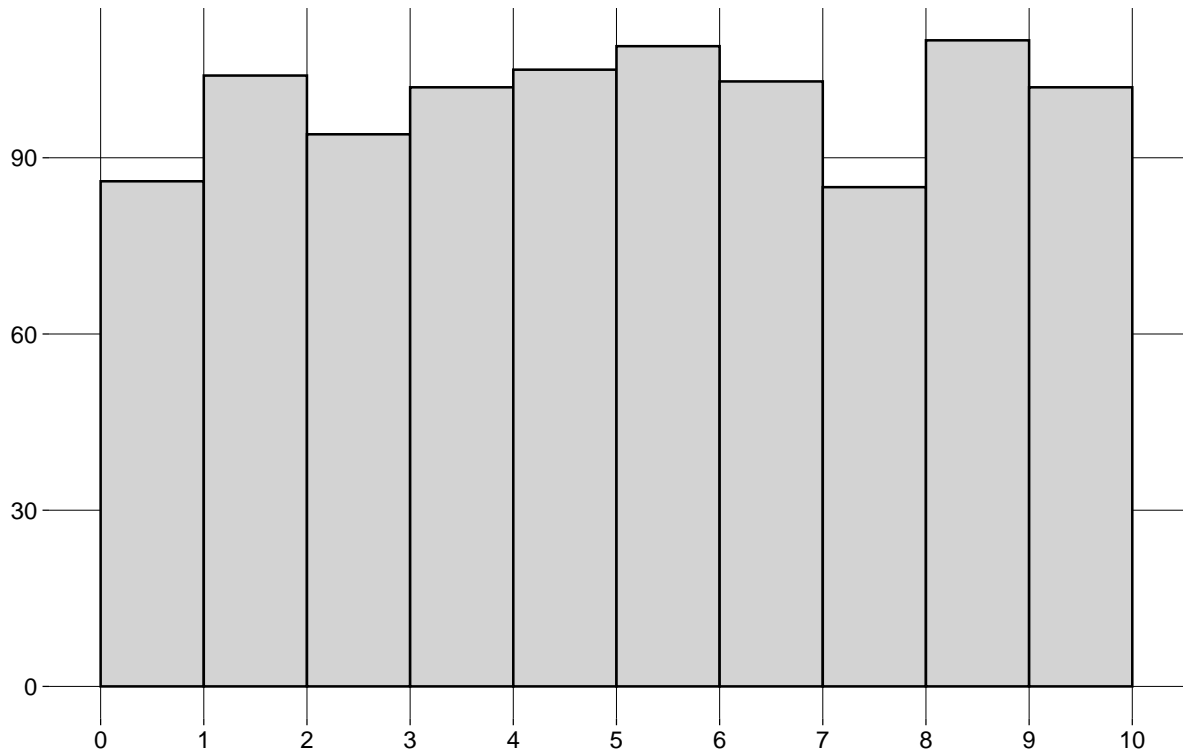
Formas de uma distribuição

- A forma do histograma mostra aspectos importantes da distribuição dos dados.

6.3.1

Distribuição Uniforme

- Se o histograma tem todas as barras aproximadamente da mesma altura, dizemos que a distribuição é **uniforme**:



- A **distribuição uniforme não tem moda**, já que todos os valores têm aproximadamente a mesma frequência.

6.3.2

Simetria

- Se o histograma for simétrico (i.e., os lados esquerdo e direito são “espelhados”), dizemos que a distribuição é **simétrica**.
- A distribuição normal **do exemplo acima** é simétrica.
- A distribuição uniforme também é simétrica.
- Para distribuições simétricas, a média, a mediana e a moda **(quando existe e é única)** são bem próximas.
 - Para a distribuição normal do exemplo:

```
mean(normal)
```

```
## [1] 4,991974
```

```
median(normal)
```

```
## [1] 4,979183
```

```
mlv(normal, method = 'venter')
```

```
## [1] 4,882353
```

– Para a distribuição uniforme do exemplo:

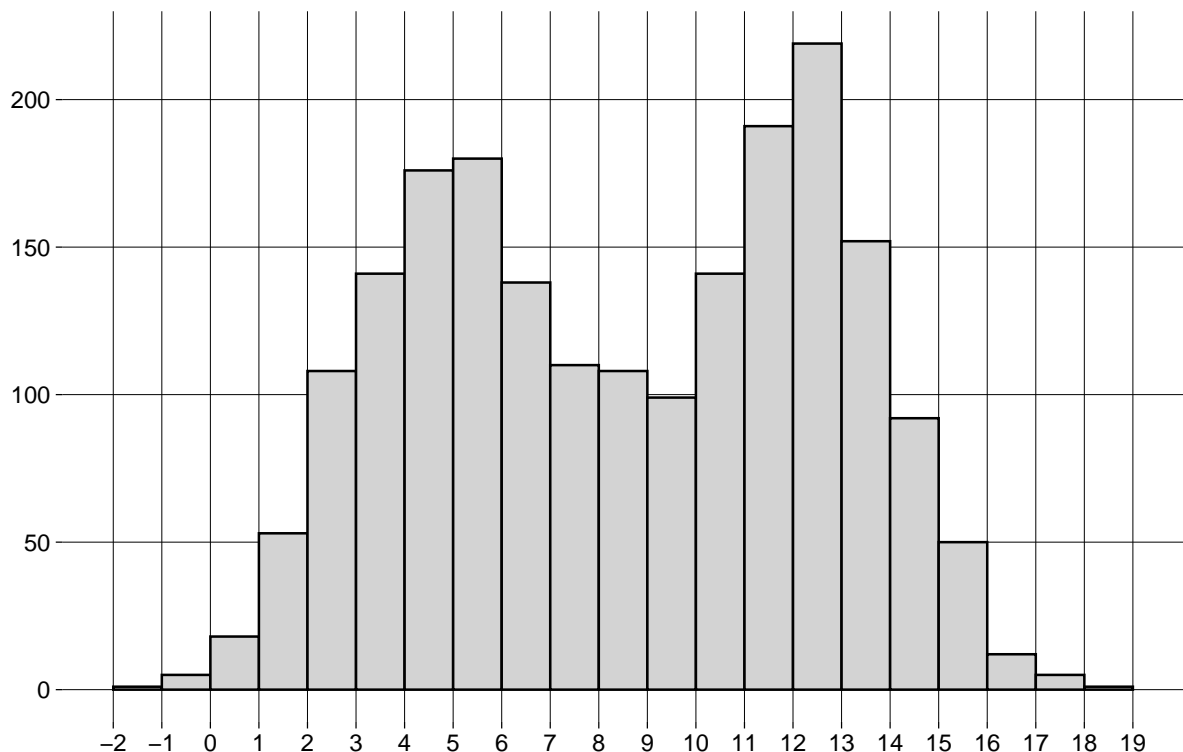
```
mean(uniforme)
```

```
## [1] 5,076746
```

```
median(uniforme)
```

```
## [1] 5,06253
```

- Uma distribuição pode ser **simétrica**, mas ter **duas (ou mais) modas diferentes**:



- Algumas distribuições não são simétricas, mas têm uma **cauda longa** à esquerda ou à direita.
- Dependendo da cauda, as distribuições são chamadas de **assimétricas à esquerda** ou **assimétricas à direita**.
- Um exemplo: receitas anuais (em milhões de dólares) de CEOs de grandes empresas:

```
df <- read_csv(
  './data/CEO_Salary_2012.csv',
  show_col_types = FALSE
)
glimpse(df)
```

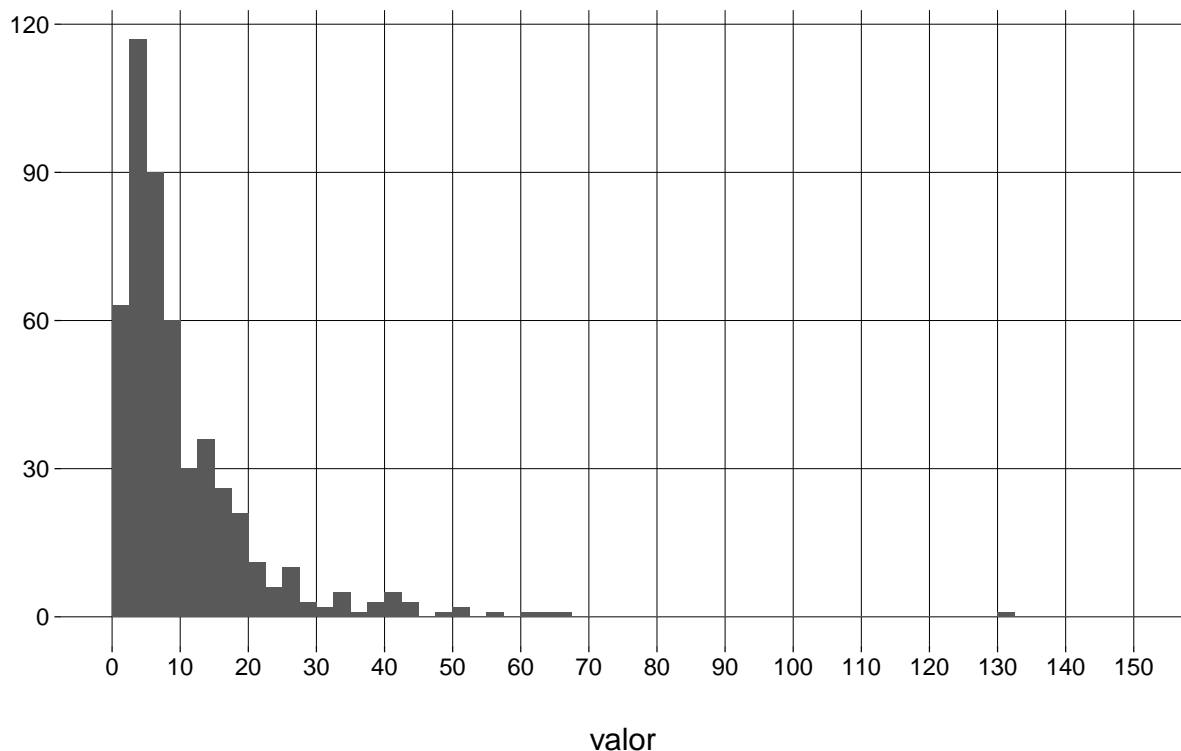
```
## Rows: 500
## Columns: 9
## $ Rank          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ~
## $ Name          <chr> "John H Hammergren", "Ralph Lauren", "M~
## $ Company       <chr> "McKesson", "Ralph Lauren", "Vornado Re~
## $ `1-Year Pay ($mil)` <dbl> 131,190, 66,650, 64,405, 60,940, 55,790~
## $ `5 Year Pay ($mil)` <dbl> 285,020, 204,060, NA, 60,940, 96,110, 1~
## $ `Shares Owned ($mil)` <dbl> 51,9, 5010,4, 171,7, 8582,3, 21,5, 47,3~
## $ Age           <dbl> 53, 72, 55, 67, 59, 57, 55, 59, 61, 60,~
## $ Efficiency    <dbl> 121, 84, NA, NA, 138, 36, 12, NA, 91, 1~
## $ `Log Pay`     <dbl> 8,117901, 7,823800, 7,808920, 7,784902,~
```

- Vamos usar apenas os nomes e os valores anuais:

```
salarios <- df %>%
  select(Name, valor = `1-Year Pay ($mil)`)
```

- Um histograma:

```
salarios %>%
  ggplot(aes(x = valor)) +
    geom_histogram(breaks = seq(0, 150, 2.5)) +
    scale_x_continuous(breaks = seq(0, 150, 10)) +
    labs(y = NULL)
```



- É uma distribuição **assimétrica à direita**: a maior parte dos CEOs têm receitas anuais “baixas”, de menos de 10 milhões. À medida que examinamos valores maiores, a quantidade de CEOs vai diminuindo lentamente.
- Observe que **a longa cauda à direita “puxa” a média para um valor mais alto do que a mediana**.
- A moda, que corresponde à barra mais alta do histograma, é menor que a mediana (e que a média):

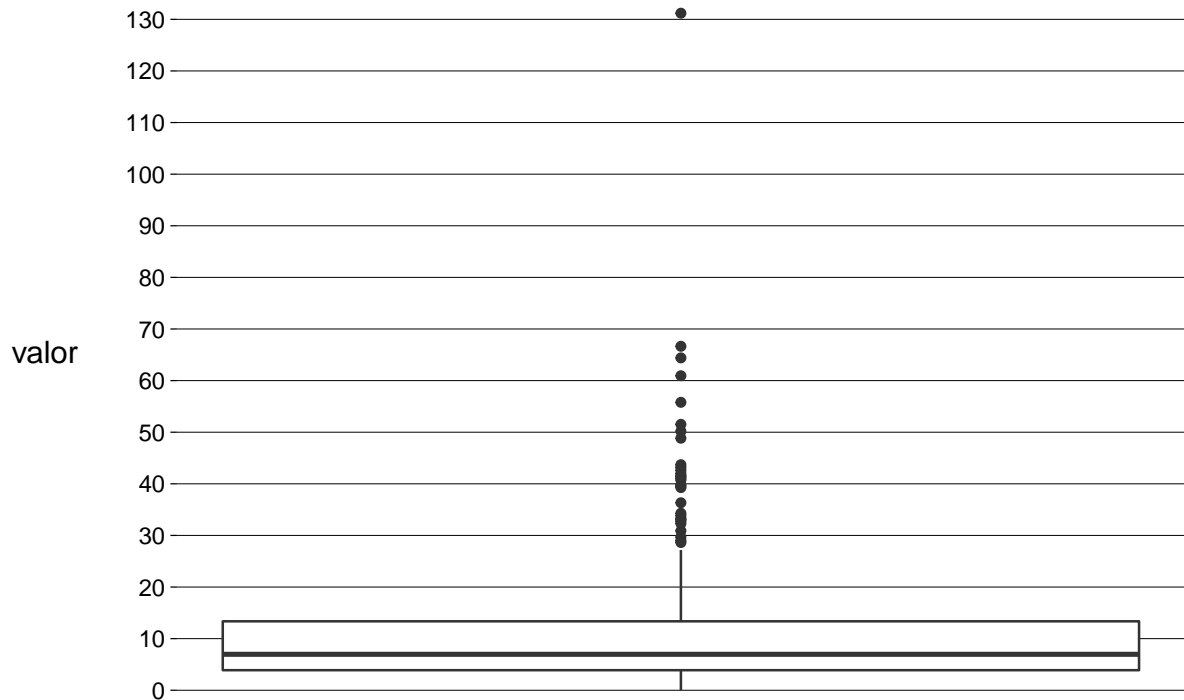
```
sumario <- salarios %>%
  summarise(
    moda = mlv(valor, method = 'venter'),
    mediana = median(valor),
    media = mean(valor)
  )

sumario
```

```
## # A tibble: 1 x 3
##   moda mediana media
##   <dbl>   <dbl> <dbl>
## 1  4.60    6.97  10.5
```

- Em um *boxplot*, também é possível detectar a assimetria pela grande quantidade de *outliers* em um extremo:


```
salarios %>%
  ggplot(aes(y = valor)) +
    geom_boxplot() +
    scale_x_continuous(breaks = NULL) +
    scale_y_continuous(breaks = seq(0, 150, 10))
```



- Com distribuições **assimétricas à esquerda**, a situação se inverte: **a média é menor que a mediana, que é menor que a moda.**

6.3.3

Exercícios

- Ache um conjunto de dados com uma distribuição assimétrica à esquerda.
- Faça um histograma.
- Calcule a média, a mediana, e a moda dos dados.

6.4

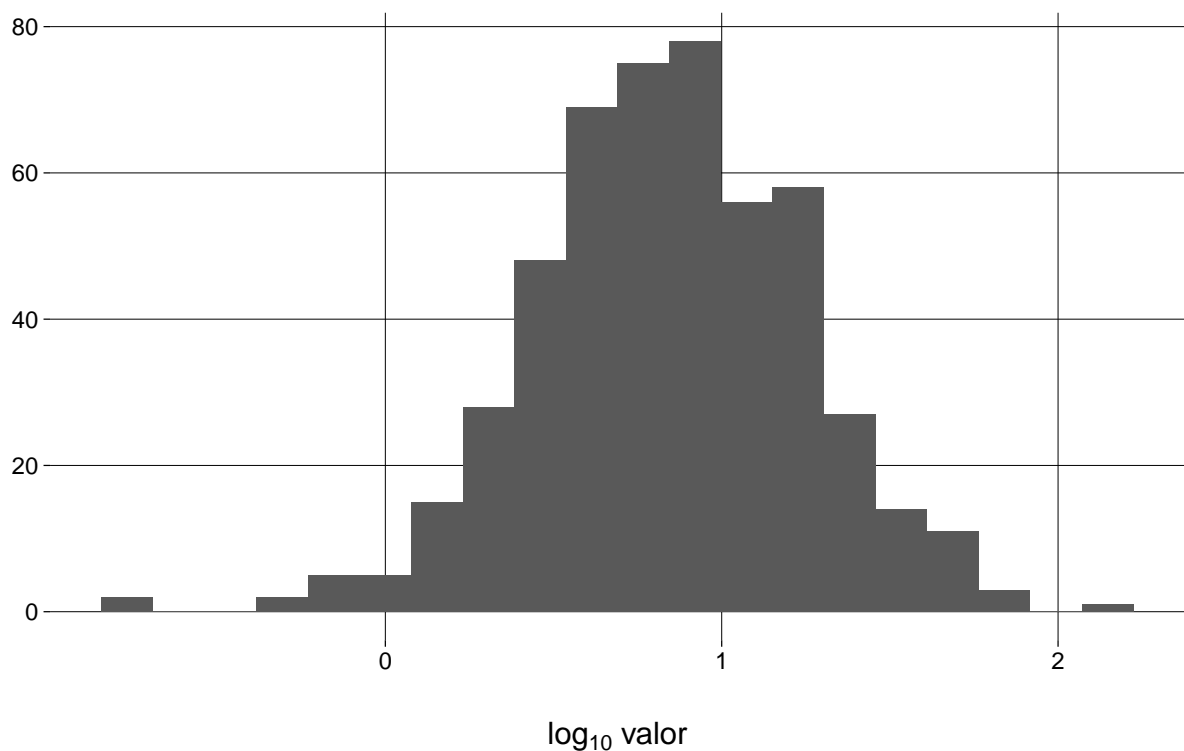
Re-expressão

- Muitas vezes, é recomendável **transformar a escala dos dados** para que uma distribuição assimétrica se torne simétrica.
- No exemplo das receitas dos CEOs, podemos tomar os **logaritmos** dos valores, em vez dos valores:

```
salarios_log <- salarios %>%  
  mutate(log_valor = log10(valor))
```

```
salarios_log %>%  
  ggplot(aes(x = log_valor)) +  
    geom_histogram(bins = 20) +  
    labs(  
      x = TeX('$\\log_{10}$ valor'),  
      y = NULL  
    )
```

Warning: Removed 3 rows containing non-finite values (stat_bin).



- O logaritmo de um número na base 10 é, essencialmente, a quantidade de dígitos do número, vista como uma grandeza contínua.
- Logaritmos negativos vêm de valores entre 0 e 1.

- Logaritmo zero vem do valor 1.
- Valores iguais ou menores que zero não têm logaritmo definido.
- Por isso a mensagem de aviso sobre 3 valores removidos. São valores iguais a zero:

```
salarios_log %>%
  filter(valor == 0)
```

```
## # A tibble: 3 x 3
##   Name          valor log_valor
##   <chr>         <dbl>    <dbl>
## 1 Malon Wilkus      0      -Inf
## 2 Matthew J Lambiase 0      -Inf
## 3 Larry Page       0      -Inf
```

- Uma vantagem desta escala logarítmica é que podemos entender melhor o histograma. Os dados não estão amontoados de um lado só.

6.4.1

Exercícios

- Quais são os registros com $\log_{10} \text{valor} < 0$?
- Faça um *boxplot* dos logaritmos das receitas.

6.5

Medidas de posição

6.5.1

Quantis

- Na *seção sobre boxplots*, falamos sobre **quantis**, que são medidas de posição.
- Em R, a função `quantile` calcula quantis de um vetor:

```
salarios %>%
  pull(valor) %>%
  quantile()
```

```
##           0%          25%          50%          75%         100%
## 0,00000  3,88500  6,96750 13,36125 131,19000
```

- Você pode passar frações entre 0 e 1 para `quantile`. Por exemplo, para calcular o primeiro, o quinto, e o décimo percentis³ das receitas dos CEOs:

³Um percentil é um quantil da forma $k/100$, para k natural, $0 \leq k \leq 100$.

```
salarios %>%
  pull(valor) %>%
  quantile(c(.01, .05, .1))
```

```
##      1%      5%     10%
## 0,48695 1,48405 2,19400
```

6.6

Medidas de dispersão

- Tão importantes quanto as medidas de centralidade são as medidas de dispersão (ou **espalhamento**).
- Elas informam **o quanto os dados variam**.

6.6.1

Amplitude

- Uma medida simples é a **diferença entre o valor máximo e o valor mínimo**.
- Lembrando do nosso exemplo das idades dos alunos:

```
idades
```

```
## [1] 20 20 20 20 20 20 20 21 21 21 21 22 22 22 23 23 23 23 24 24 65
```

- A função `range` retorna o mínimo e o máximo:

```
range(idades)
```

```
## [1] 20 65
```

- A amplitude destes dados é, então

```
range(idades)[2] - range(idades)[1]
```

```
## [1] 45
```

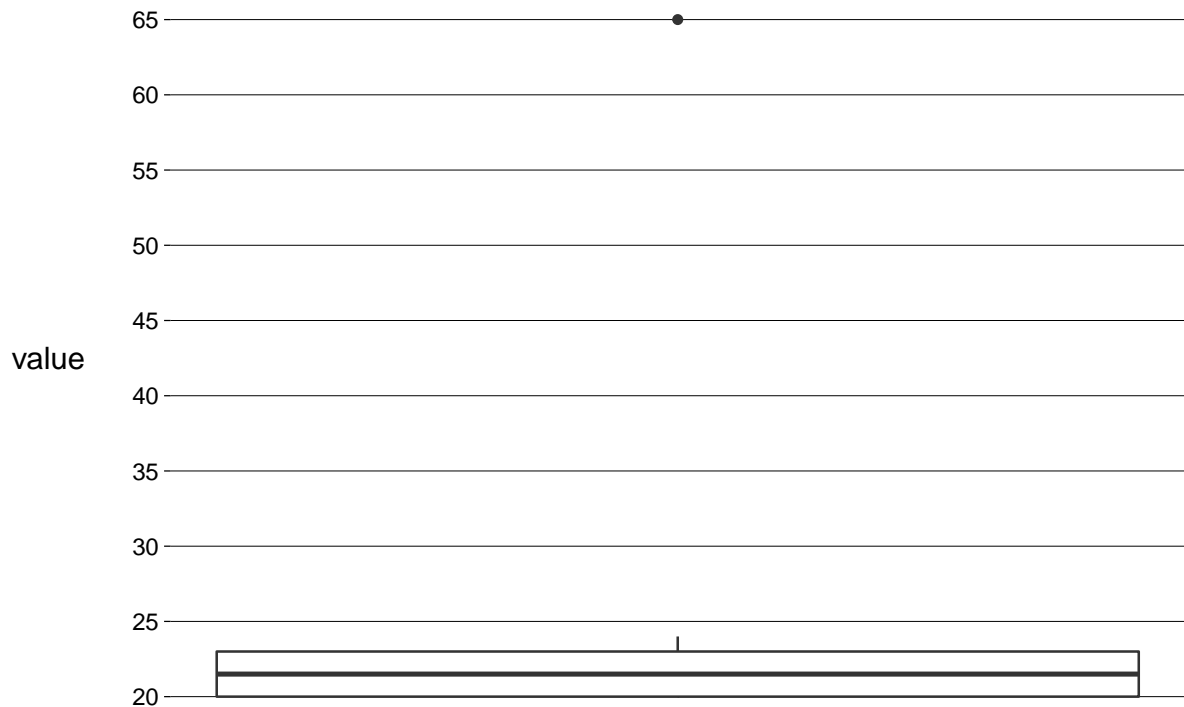
- A diferença de idade entre o aluno mais novo e o mais velho é de 45 anos, um valor alto, por causa do velhinho.

6.6.2

IQR

- Na *seção sobre boxplots*, também falamos sobre o **intervalo interquartil** (IQR).
- No *boxplot*, é a **altura da caixa**. Para as idades dos alunos:

```
idades %>%  
  as_tibble() %>%  
  ggplot(aes(y = value)) +  
    geom_boxplot() +  
    scale_x_continuous(breaks = NULL) +  
    scale_y_continuous(breaks = seq(20, 70, 5))
```



- O IQR é a diferença entre o primeiro e o terceiro quartil:

```
summary(idades)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
##  20,00  20,00   21,50   23,75  23,00   65,00
```

```
unname(summary(idades)[5] - summary(idades)[2])
```

```
## [1] 3
```

```
IQR(idades)
```

```
## [1] 3
```

- Ou seja, os 50% centrais dos alunos têm idade entre 20 e 23 anos, um IQR de 3.
- É uma variação pequena, porém mais fiel à realidade do que a amplitude, que é alta por causa do velhinho.
- Quanto maior o IQR, mais espalhados estão os dados.

6.6.3

Variância

- Agora, vamos trabalhar com os pesos (kg) e alturas (m) de um time de basquete:

```
medidas <- tibble(  
  altura = .025 *  
    c(72, 74, 68, 76, 74, 69, 72, 79, 70, 69, 77, 73),  
  peso = 0.45 *  
    c(180, 168, 225, 201, 189, 192, 197, 162, 174, 171, 185, 210)  
)  
  
medidas
```

```
## # A tibble: 12 x 2  
##   altura peso  
##   <dbl> <dbl>  
## 1   1.8   81  
## 2   1.85  75.6  
## 3   1.7  101.  
## 4   1.9   90.4  
## 5   1.85  85.0  
## 6   1.72  86.4  
## # ... with 6 more rows
```

```
summary(medidas$altura)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
##  1,700  1,744   1,812   1,819  1,863   1,975
```

```
summary(medidas$peso)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
##  72,90  77,96   84,15   84,53  89,10  101,25
```

- A variância é a maneira mais usada de medir o espalhamento em torno da média.

- Para calcular a variância das alturas e a variância dos pesos, precisamos calcular valores intermediários.
- O **desvio** de um valor é a **diferença entre o valor e a média**. O desvio pode ser positivo ou negativo.

```
d_medidas <- medidas %>%
  mutate(
    d_altura = altura - mean(altura),
    d_peso = peso - mean(peso)
  )

d_medidas
```

```
## # A tibble: 12 x 4
##   altura peso d_altura d_peso
##   <dbl> <dbl>   <dbl>  <dbl>
## 1   1.8   81    -0.0188 -3.53
## 2   1.85  75.6    0.0312 -8.92
## 3   1.7  101.   -0.119  16.7
## 4   1.9   90.4    0.0813  5.92
## 5   1.85  85.0    0.0312  0.525
## 6   1.72  86.4   -0.0938  1.88
## # ... with 6 more rows
```

- Vamos calcular o desvio médio das alturas e o desvio médio dos pesos:

```
d_medidas %>%
  summarize(
    d_medio_altura = mean(d_altura),
    d_medio_peso = mean(d_peso)
  )
```

```
## # A tibble: 1 x 2
##   d_medio_altura d_medio_peso
##           <dbl>         <dbl>
## 1             0    -3.55e-15
```

- Não foi uma boa idéia. **O desvio médio sempre é igual a zero.**⁴ (O R pode mostrar algum valor diferente de zero por causa da precisão limitada dos números de ponto flutuante.)
- Como resolver isto? Elevando os desvios ao quadrado:

```
dq_medidas <- d_medidas %>%
  mutate(
    dq_altura = d_altura^2,
```

⁴Você vai provar isto em um exercício.

```

    dq_peso = d_peso^2
  )

```

```

dq_medidas

```

```

## # A tibble: 12 x 6
##   altura peso d_altura d_peso dq_altura dq_peso
##   <dbl> <dbl>   <dbl>  <dbl>   <dbl>   <dbl>
## 1   1.8   81    -0.0188 -3.53   0.000352  12.4
## 2   1.85  75.6    0.0312 -8.92   0.000977  79.7
## 3   1.7  101.   -0.119  16.7    0.0141   280.
## 4   1.9   90.4    0.0813  5.92    0.00660   35.1
## 5   1.85  85.0    0.0312  0.525   0.000977   0.276
## 6   1.72  86.4   -0.0938  1.88    0.00879   3.52
## # ... with 6 more rows

```

- Agora temos os **desvios quadrados**, que são todos **positivos**.
- O **desvio quadrado médio** vai ser a **variância**:

```

dq_medidas %>%
  summarize(
    var_altura = mean(dq_altura),
    var_peso = mean(dq_peso)
  )

```

```

## # A tibble: 1 x 2
##   var_altura var_peso
##   <dbl>     <dbl>
## 1   0.00678     63.3

```

- Uma vantagem da variância é que *outliers* (que têm desvios quadrados maiores) contribuem mais do que elementos próximos à média (que têm desvios quadrados menores).
- Uma desvantagem da variância é que a **sua unidade é o quadrado da unidade dos valores**.
- Neste exemplo, as unidades são m^2 e kg^2 !

6.6.4

Desvio-padrão

- É melhor trabalhar com **a raiz quadrada da variância**, que chamamos de **desvio-padrão**.
- As unidades são as mesmas que as unidades dos dados.


```
dq_medidas %>%
  summarize(
    dp_altura = sqrt(mean(dq_altura)),
    dp_peso = sqrt((mean(dq_peso)))
  )
```

```
## # A tibble: 1 x 2
##   dp_altura dp_peso
##   <dbl>    <dbl>
## 1    0.0824    7.96
```

- Claro que o R tem funções para calcular isso: `var` e `sd` (*standard deviation*):

```
medidas %>%
  summarize(
    altura_var = var(altura),
    altura_dp = sd(altura),
    peso_var = var(peso),
    peso_dp = sd(peso)
  )
```

```
## # A tibble: 1 x 4
##   altura_var altura_dp peso_var peso_dp
##   <dbl>    <dbl>    <dbl>    <dbl>
## 1    0.00740    0.0860    69.1    8.31
```

- Mas os valores são diferentes dos que calculamos. Por quê?

6.6.5

Definições

- Para uma **população** com N elementos e média μ , a **variância** é

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

- e o **desvio-padrão** é

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

- Para uma **amostra** com n elementos e média \bar{x} , a **variância** é

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

e o desvio-padrão é

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

- Nós calculamos a versão populacional destas medidas.
- R calcula a versão amostral destas medidas.
- Reveja os cálculos e entenda a diferença.
- Note, também, que as medidas populacionais são representadas por letras gregas — μ, σ^2, σ —, enquanto as medidas amostrais são representadas por letras latinas — \bar{x}, s^2, s .



Mais adiante no curso, você vai entender por que o denominador da variância amostral é $n - 1$, em vez de n .

Nada é por acaso, nem mesmo em Estatística.

6.6.6

Exercícios

- Quando a variância e o desvio-padrão de um conjunto de dados são iguais a zero?
- Mostre que o desvio médio de qualquer conjunto de valores é igual a zero.

Ou seja, considere o conjunto

$$\{x_1, x_2, \dots, x_n\}$$

e prove que

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

Manipule apenas as variáveis x_i . Não use exemplos, pois eles não provam o enunciado geral.

Dica: lembre que $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$.

Coeficiente de variação

- Em um conjunto de dados, o desvio-padrão é uma medida importante da variação dos dados.
- Mas a unidade do desvio-padrão muda de um conjunto de dados para outro: alturas em metros, pesos em quilos etc.
- Podemos eliminar as unidades expressando o desvio-padrão em termos da média.
- O resultado é a fração $\frac{\sigma}{\mu}$ (na população) ou $\frac{s}{\bar{x}}$ na amostra.
- Esta fração é o coeficiente de variação (CV).
- O CV não tem unidades.
- Para as alturas do exemplo dos jogadores de basquete:

A média das alturas é 1,82 metros.

O desvio-padrão das alturas é 0,09 metros.

O CV é aproximadamente 0,0473.

```
statip::cv(medidas$altura)
```

```
## [1] 0,04729982
```

Em outras palavras, para as alturas, um desvio-padrão corresponde a 4,73% da média.

- Para os pesos:

A média dos pesos é 84,53 quilos.

O desvio-padrão dos pesos é 8,31 quilos.

O CV é aproximadamente 0,0983.

```
statip::cv(medidas$peso)
```

```
## [1] 0,09834649
```

Em outras palavras, para os pesos, um desvio-padrão corresponde a 9,83% da média.

- Segundo estes valores, a variação dos pesos é cerca de 2 vezes maior do que a variação das alturas.



O coeficiente de variação sempre faz sentido para dados do **nível racional** (veja a definição) — i.e., dados onde o zero é absoluto.

Para dados apenas intervalares, o uso do CV pode levar a conclusões absurdas, como você terá chance de ver no exercício.

6.7.1

Exercícios

- Considere o seguinte conjunto de temperaturas (em graus Celsius):

```
celsius <- c(0, 10, 20, 30, 40)
```

- E as **mesmas temperaturas** (em graus Fahrenheit):

```
fahrenheit <- 9 * celsius / 5 + 32
```

- Calcule para cada um dos dois vetores acima:
 1. A média,
 2. O desvio-padrão,
 3. O coeficiente de variação.
- As temperaturas são as mesmas (apenas em unidades diferentes), mas os CVs são diferentes. Por quê?
- Agora, convertamos as **mesmas temperaturas** para a **escala Kelvin**:

```
kelvin <- celsius + 273.15
```

- E para a **escala Rankine**:

```
rankine <- fahrenheit + 459.67
```

- Calcule para cada um dos dois vetores acima:
 1. A média,
 2. O desvio-padrão,
 3. O coeficiente de variação.
- Compare:
 1. As médias de `celsius` e `kelvin`,
 2. As médias de `fahrenheit` e `rankine`,
 3. Os desvios-padrão de `celsius` e `kelvin`,

4. Os desvios-padrão de fahrenheit e rankine,
 5. Os coeficientes de variação de kelvin e rankine.
- Explique o que houve.

6.8

Escores-padrão

- Para qualquer conjunto de dados, a unidade usada é uma escolha arbitrária.
- Para alturas, por exemplo, podemos usar metros, centímetros, pés, polegadas etc.
- A escolha de unidades é tão arbitrária que podemos escolher uma unidade (que dificilmente vai ter nome) que faça com que a média do conjunto de dados seja zero e que o desvio-padrão seja igual a 1.
- Isto equivale a tomar, como unidade, o *desvio-padrão acima da média*.
- Os valores, nesta nova unidade, são chamados de *escores-padrão*.
- Dizemos que os valores foram *padronizados*.
- Vamos usar as alturas dos jogadores de basquete.
- Para fazer a altura média virar zero, basta subtrair, de cada altura, a altura média:

```
alturas <- medidas$altura
mean(alturas)
```

```
## [1] 1,81875
```

```
alturas_deslocadas <- alturas - mean(alturas)
mean(alturas_deslocadas)
```

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

- Para fazer o desvio-padrão ser igual a 1, basta dividir estes valores pelo desvio-padrão dos dados originais:

```
sd(alturas)
```

```
## [1] 0,08602656
```

```
alturas_padronizadas <- alturas_deslocadas / sd(alturas)
sd(alturas_padronizadas)
```

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

- Tome, por exemplo, o seguinte jogador:

```
altura <- alturas[1]
altura
```

```
## [1] 1,8
```

```
altura_padronizada <- (alturas[1] - mean(alturas)) / sd(alturas)
altura_padronizada
```

```
## [1] -0,217956
```

Faça as contas: o valor da altura padronizada deste jogador significa que a altura dele está 0,217956 **desvios-padrão abaixo da altura média**.

- **No geral:**

- Se a média for \bar{x} , e
- Se o desvio-padrão for s ,
- Os escores-padrão z_i vão ser

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Em R, a função `scale` faz isso:

```
medidas <- medidas %>%
  mutate(altura_padronizada = scale(altura)[,1])

medidas %>%
  select(altura, altura_padronizada)
```

```
## # A tibble: 12 x 2
##   altura altura_padronizada
##   <dbl>         <dbl>
## 1   1.8         -0.218
## 2   1.85         0.363
## 3   1.7        -1.38
## 4   1.9         0.944
## 5   1.85         0.363
## 6   1.72        -1.09
## # ... with 6 more rows
```

```
mean(medidas$altura_padronizada)
```

```
## [1] -0,000000000000000004610683
```

```
sd(medidas$altura_padronizada)
```

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

- A função `scale` foi feita para receber e retornar **matrizes**. Como estamos trabalhando com **vetores**, usamos `scale(altura)[,1]` para tomar apenas a primeira (e única) coluna do resultado.

6.8.1

Exercícios

- Por que, **quando calculamos as alturas deslocadas divididas pelo desvio-padrão das alturas**, temos certeza de que a média dos valores resultantes não mudou?
- Padronize os pesos dos jogadores de basquete.
- Confira a média e o desvio-padrão dos pesos padronizados.
- Crie um *scatterplot* de peso por altura.
- Crie um *scatterplot* de peso padronizado por altura padronizada.
- Compare os dois *scatterplots*. O que muda de um para outro?

6.9

Teorema de Tchebychev

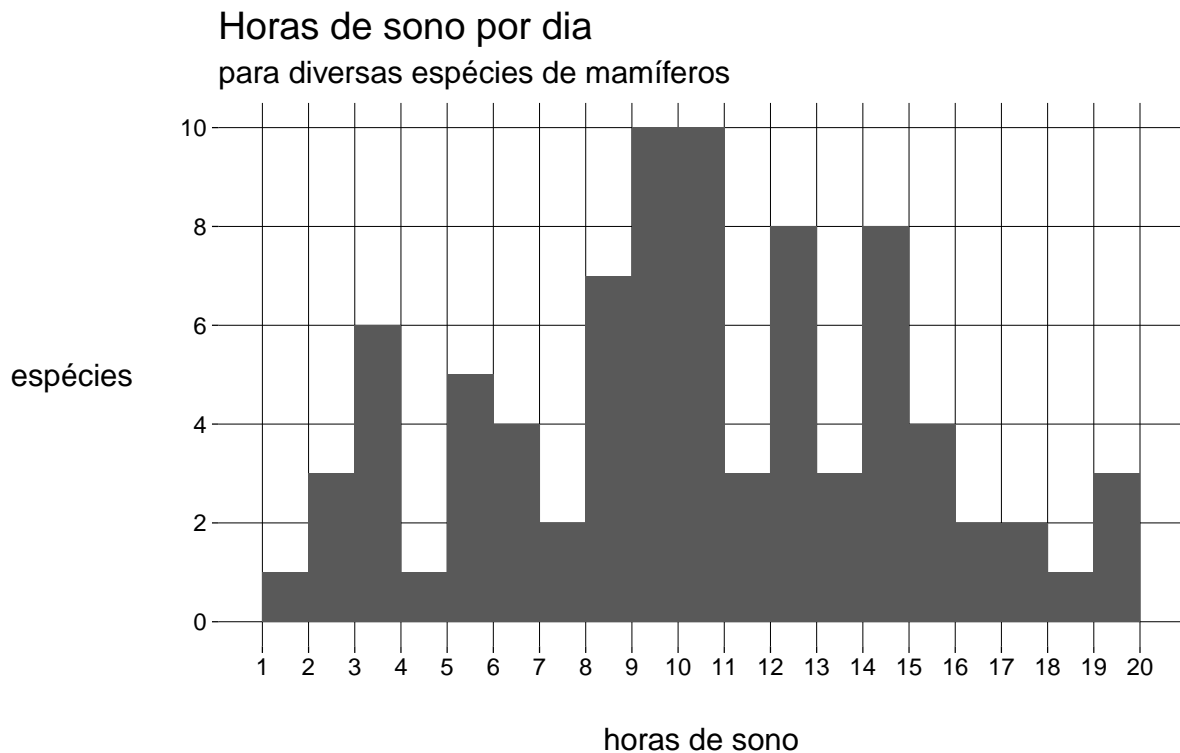
- Grosso modo, **quanto mais alto o desvio-padrão, maior é a distância da média até os valores**.
- Ou seja, **quanto menor o desvio-padrão, maior é a proporção de valores que estão próximos à média**.
- O teorema de Tchebychev quantifica esta idéia:
Em **qualquer** distribuição, a proporção de valores **dentro de $\pm k$ desvios-padrão ($k > 1$) da média** é de, **no mínimo**

$$1 - \frac{1}{k^2}$$

6.9.1

Exemplo

- Lembre-se do **conjunto de dados sobre os totais de horas de sono de diversos mamíferos**:



- Média e desvio-padrão:

```
media <- mean(df$value)
media
```

```
## [1] 10,43373
```

```
dp <- sd(df$value)
dp
```

```
## [1] 4,450357
```

- Qual a proporção de espécies que estão a 1,3 ou menos desvios-padrão de distância da média?

```
k <- 1.3
inicio <- media - k * dp
inicio
```

```
## [1] 4,648271
```

```
fim <- media + k * dp
fim
```

```
## [1] 16,2192
```

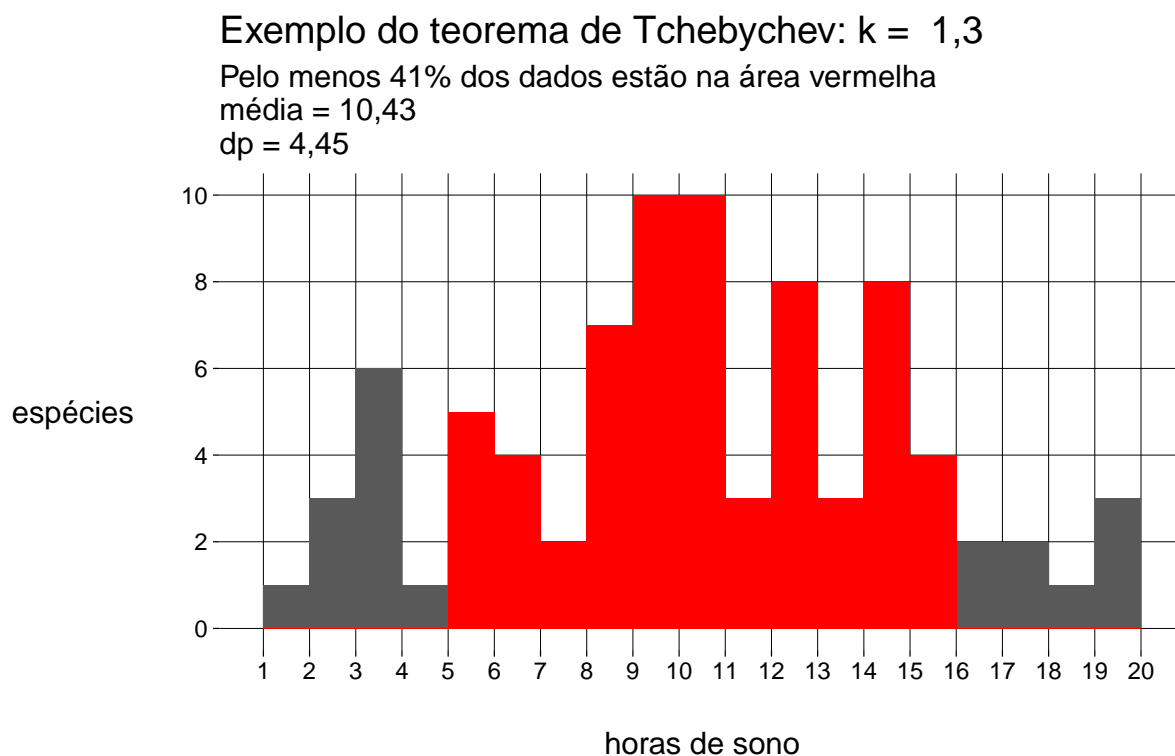
- O teorema diz que no mínimo a seguinte proporção das espécies está dentro deste

intervalo:

```
proporcao_teorema <- 1 - 1 / k^2  
proporcao_teorema
```

```
## [1] 0,408284
```

- Graficamente:



- Vamos conferir:

```
total_especies <- df %>% nrow()  
total_especies
```

```
## [1] 83
```

```
especies_intervalo <- df %>%  
  filter(value >= inicio & value <= fim) %>%  
  nrow()  
especies_intervalo
```

```
## [1] 64
```

```
proporcao_real <- especies_intervalo / total_especies  
proporcao_real
```

[1] 0,7710843

- Como o teorema usa apenas a média e o desvio-padrão, e **mais nenhuma informação sobre a distribuição dos valores** — forma, simetria etc. — o que ele garante é, muitas vezes, mais fraco do que a realidade.
- Neste exemplo, o teorema garantia **no mínimo** 40,83% das espécies a 1,30 ou menos desvios-padrão de distância da média.
- A proporção verdadeira é 77,11% das espécies.
- O teorema está certo (claro), mas, sem mais informações sobre a distribuição dos dados, o teorema não pode ser mais preciso.

6.9.2

Exercício

- Uma loja recebe uma média de 20 clientes por dia, com desvio-padrão de 2 clientes.
- Em um ano (365 dias), em quantos dias, no mínimo, o número de clientes ficou entre 16 e 24, inclusive?

Probabilidades

7.1

Vídeo 1

<https://youtu.be/hJfyzRzEs44>

7.2

Espaço amostral

- Para falar em probabilidades, precisamos falar de experimentos, resultados, espaços amostrais, e eventos.
- Um experimento probabilístico é um experimento cujo resultado exato é desconhecido *a priori*; mais ainda: executar o experimento diversas vezes, nas mesmas condições, pode produzir resultados diferentes.
- O espaço amostral é o conjunto de todos os resultados possíveis de um experimento probabilístico, representados de alguma forma.
- Exemplos:
 1. Experimento: lançar uma moeda;
Espaço amostral: $\{K, C\}$ (onde K é cara, C é coroa).

2. Experimento: lançar 2 moedas;
Espaço amostral: $\{(K, K), (K, C), (C, K), (C, C)\}$.
3. Experimento: lançar um dado;
Espaço amostral: $\{1, 2, 3, 4, 5, 6\}$.
4. Experimento: lançar 2 dados;
Espaço amostral: $\{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}$.

7.3

Evento

- Um **evento** é um subconjunto do espaço amostral; ou seja, um evento é um **conjunto de resultados**.
- Exemplos:
 1. Lançar uma moeda e **obter cara**:
 $\{K\}$.
 2. Lançar 2 moedas e **obter resultados iguais**:
 $\{(K, K), (C, C)\}$.
 3. Lançar um dado e **obter um número maior que 4**:
 $\{5, 6\}$.
 4. Lançar 2 dados e **obter 2 números iguais**:
 $\{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$.
- Dizemos que **o evento A ocorreu** se o experimento foi realizado e **o resultado obtido está no conjunto que corresponde ao evento A** .

7.4

Análise Combinatória

- Para calcularmos probabilidades, vamos precisar **contar a quantidade de certos objetos complexos** (formados por partes menores).
- Existem **técnicas de contagem**, que são assunto de **Análise Combinatória**.
- Exemplos:
 1. Quantas senhas de 6 caracteres (dentre letras e dígitos apenas) existem, **sem distinguir entre minúsculas e maiúsculas**?
 2. E se não puder haver repetição de caracteres?
- Se você nunca tiver estudado técnicas de contagem, ou se quiser revisar ou aprender mais, consulte o excelente livro Oliveira Morgado et al. (2004).

7.5

Probabilidade clássica

- Nesta abordagem simples — que pode não ser a correta para o problema que estamos tentando resolver —, **cada resultado do espaço amostral tem a mesma chance de ocorrer.**
- Ou seja, para um evento A , a probabilidade $P(A)$ é

$$P(A) = \frac{\text{Qtde de resultados em } A}{\text{Qtde de resultados no espaço amostral}}$$

- Exemplo: de um baralho normal, de 52 cartas, qual a probabilidade de escolher uma carta ao acaso e obter
 1. Uma carta de ouros?
 2. Uma carta vermelha?
 3. Uma carta de figura (J, Q ou K)?
 4. Uma carta de ouros, copas, paus ou espadas?
 5. Um carta de um naipe verde?



Só podemos usar este raciocínio **se todos os resultados do experimento tiverem a mesma probabilidade** de ocorrer.

Como a carta é escolhida **ao acaso**, esta condição é satisfeita neste exemplo.

7.6

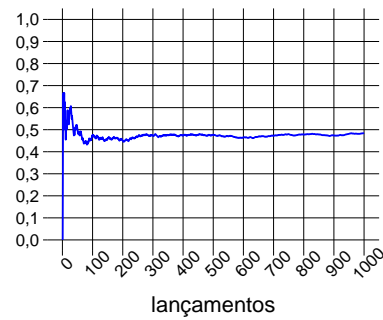
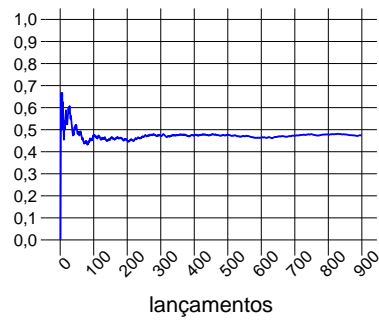
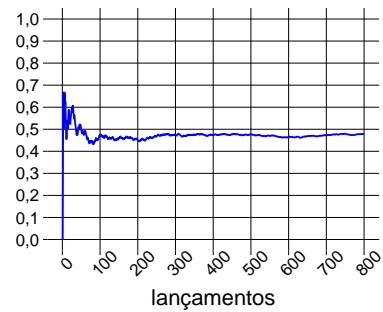
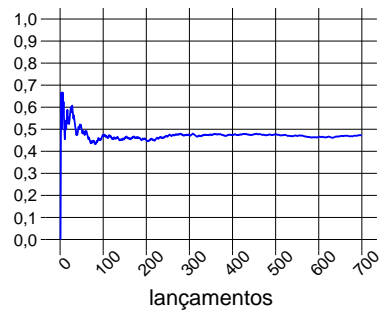
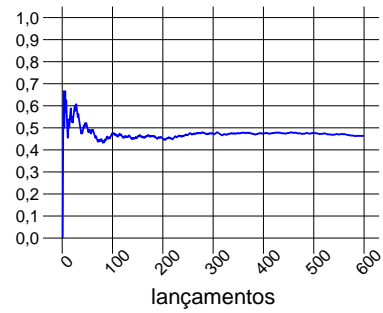
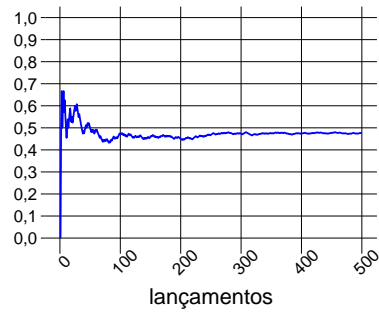
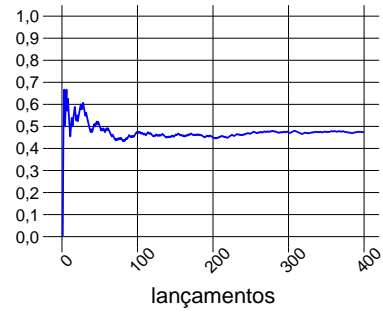
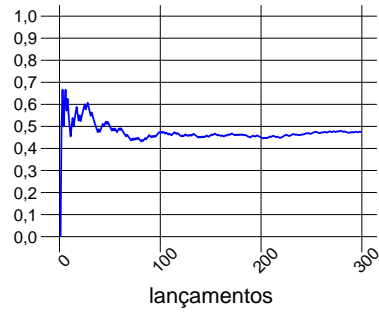
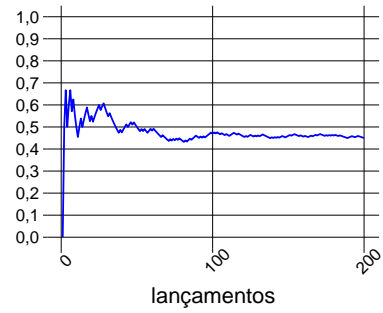
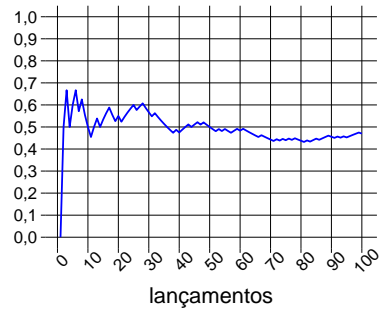
Probabilidade empírica

- Baseada em **repetições de um experimento probabilístico.**
- Nesta abordagem, a probabilidade de um evento é sua **frequência relativa:**

$$P(A) = \frac{\text{Qtde de ocorrências de } A}{\text{Qtde total de repetições do experimento}}$$

- Esta abordagem é fácil de usar quando é possível repetir o experimento muitas vezes, nas mesmas condições (lançar uma moeda, escolher uma carta de um baralho).
- Em outros casos (calcular a probabilidade de um candidato vencer uma eleição), não é possível repetir o experimento nas mesmas condições.

- Exemplo: se lançarmos uma moeda não-viciada muitas vezes, a proporção de caras vai ser aproximadamente 0,5. Os gráficos abaixo mostram como, à medida que o número de lançamentos aumenta (no eixo horizontal), a proporção de caras (no eixo vertical) vai se aproximando de 0,5:



- A **lei dos grandes números** é um resultado matemático que diz, essencialmente, que, quando o número n de repetições de um experimento tende a infinito, a frequência relativa de um evento tende à sua probabilidade real.

Um **erro comum** é achar que, se houve poucas caras nos lançamentos mais recentes, então a probabilidade de o resultado ser cara no próximo lançamento é maior, para que a proporção de caras fique mais perto de 0,5.

A lei dos grandes números fala sobre os resultados do experimento quando n tende ao infinito, **não no futuro próximo**.

Em lançamentos **independentes** de uma moeda **não-viciada**, a probabilidade de cara **sempre** é 0,5.

7.7

Probabilidade subjetiva

- Outra interpretação de probabilidades se baseia na **crença** — a estimativa de um agente sobre a ocorrência de um evento.
- Uma maneira de quantificar a crença é através de **apostas justas**.
- Por exemplo, você aposta com um amigo que
 - Se o seu time de basquete¹ vencer o próximo jogo contra o dele, você pagará \$3 para ele.
 - Se o time dele vencer o próximo jogo contra o seu, ele pagará \$1 para você.
- Se você considera justa esta aposta, então **você crê que a probabilidade de o time dele vencer é 3 vezes maior do que a probabilidade de o seu time vencer**.
- Como **a soma das probabilidades de um evento e do evento complementar deve ser 1**, isto equivale a dizer que

$$P(\text{seu time vencer}) = 1/4 \quad \text{e} \quad P(\text{time dele vencer}) = 3/4$$

- Em mais detalhes:
 - Você pode receber menos com uma probabilidade maior,
 - Seu amigo pode receber mais com uma probabilidade menor,
 - A razão entre as quantias (3) é contrabalançada exatamente pela razão entre as probabilidades (1/3).

¹Considere que, no basquete, o empate é impossível.

7.8

Formalização de probabilidades

- Para trabalhar matematicamente com probabilidades, é preciso definir as “regras do jogo”.
- **Tudo** que se pode concluir sobre probabilidades é **consequência dos seguintes axiomas, formulados por Kolmogorov em 1933:**
 1. $0 \leq P(A) \leq 1$, para qualquer evento A .
 2. $P(\Omega) = 1$, onde Ω é o espaço amostral (o conjunto de todos os resultados possíveis do experimento em questão);
 3. $P(\bar{A}) = 1 - P(A)$, onde \bar{A} é o evento **complementar** de A (i.e., o evento que corresponde a A **não** ocorrer)
 4. $P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n)$, onde A_1, \dots, A_n são eventos **dis-juntos dois a dois** (i.e., A_i e A_j **não podem ocorrer ao mesmo tempo**, para todo par (i, j) com $i \neq j$).
- Mostre, a partir dos axiomas acima, que

$$P(\emptyset) = 0$$

Dica

$$\emptyset = \bar{\Omega}$$

7.9

Eventos independentes (explicação informal)

- Se a ocorrência de A **não influencia** a ocorrência de B , nem vice-versa, dizemos que os eventos A e B são **independentes**.
- Exemplo:
 - O experimento é **lançar um dado duas vezes**.
 - A é o evento o **primeiro lançamento deu um número par**.
 - B é o evento **o segundo lançamento deu 6**.
 - Saber se A aconteceu **não nos ajuda** em nada a estimar se B aconteceu.
 - Aqui, **A e B são independentes**.

- Outro exemplo:
 - O experimento é **lançar um dado duas vezes**.
 - A é o evento o **primeiro lançamento deu um número menor que 3**.
 - B é o evento **a soma dos dois lançamentos é maior que 8**.
 - Agora, saber se A aconteceu **ajuda** a estimar se B aconteceu.
 - Na verdade, **se A aconteceu, B é impossível** (a probabilidade de B , dado A , é 0).
 - Se A não aconteceu, a probabilidade de B é $5/12$.
 - Aqui, A e B **não são independentes**.
- A probabilidade de os eventos A e B acontecerem ao mesmo tempo é escrita como

$$P(A, B) \quad \text{ou como} \quad P(A \cap B)$$

- Quando A e B são **independentes**,

$$P(A, B) = P(A) \cdot P(B)$$

- Ou seja, **quando A e B são independentes, a probabilidade de A e B acontecerem ao mesmo tempo é igual ao produto das probabilidades de A e de B** .
- Mais adiante, vamos ver uma definição formal de independência, e vamos provar esta última igualdade.

7.10

$P(A \cup B)$ com A e B não-disjuntos

- Um dos **axiomas de probabilidade** fala sobre a probabilidade da união de vários eventos **disjuntos** (sem elementos em comum):

$$P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n)$$

- E **se os eventos não forem disjuntos?**
- Veja a figura abaixo:
???
- Imagine que a probabilidade de um evento é proporcional à sua área nesta figura.
- Se você somar a área de A com a área de B , você vai estar **contando duas vezes** a área comum aos dois (a área que corresponde a $A \cap B$).
- Por isso, **o certo é “descontar” esta área**.

- O resultado é

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Exemplo: suponha que 25% das pessoas têm cachorro, 29% das pessoas têm gato, e 12% das pessoas têm cachorro e gato.
- Qual a probabilidade de que uma pessoa tenha gato ou cachorro ou ambos?

$$\begin{aligned} P(\text{cachorro} \cup \text{gato}) &= P(\text{cachorro}) + P(\text{gato}) - P(\text{cachorro} \cap \text{gato}) \\ &= 0,25 + 0,29 - 0,12 \\ &= 0,42 \end{aligned}$$

- No geral, para n eventos A_1, \dots, A_n :

$$\begin{aligned} P(A_1 \cup \dots \cup A_n) &= P(A_1) + \dots + P(A_n) \\ &\quad - P(A_1 \cap A_2) - \dots - P(A_{n-1} \cap A_n) \\ &\quad + P(A_1 \cap A_2 \cap A_3) + \dots + P(A_{n-2} \cap A_{n-1} \cap A_n) \\ &\quad \dots \\ &\quad \pm P(A_1 \cap \dots \cap A_n) \end{aligned}$$

- Na última linha
 - o sinal vai ser $+$ se n for ímpar;
 - o sinal vai ser $-$ se n for par;
 - poderíamos escrever, então, $(-1)^{n+1} \cdot P(A_1 \cap \dots \cap A_n)$.
- Escreva, seguindo o padrão acima, a expressão para

$$P(A \cup B \cup C)$$

7.11 _____

Problema do aniversário

7.11.1 _____

Solução teórica

- Em uma sala estão 25 pessoas escolhidas ao acaso.
- Qual a probabilidade de que pelo menos 2 delas façam aniversário no mesmo dia do ano?
- Premissas:

- Os dias dos aniversários das pessoas são independentes.
- Cada dia do ano tem a mesma probabilidade de ser o aniversário de alguém.
- Vamos ignorar anos bissextos. Cada ano tem 365 dias.
- Queremos achar $P(I)$, onde I é o evento de que pelo menos duas pessoas têm aniversários iguais.
- Vamos calcular a probabilidade $P(N)$ de que **não** haja aniversários iguais.
- Este evento N é o complementar do evento I , i.e., $N = \bar{I}$.
- Então, $P(I) = 1 - P(N)$.
- $P(N)$ é a probabilidade de que todos os aniversários caiam em dias diferentes:
 - A pessoa 1 pode ter nascido em qualquer dia do ano.
 - A pessoa 2 precisa ter nascido em algum dos outros 364 dias. A probabilidade é $\frac{364}{365}$.
 - A pessoa 3 precisa ter nascido em algum dos outros 363 dias. A probabilidade é $\frac{363}{365}$.
 - ...
 - A pessoa 25 precisa ter nascido em algum dos outros 341 dias. A probabilidade é $\frac{341}{365}$.
- Como os nascimentos são independentes, a probabilidade de todos os eventos acontecerem juntos é o produto das probabilidades:

$$P(N) = \frac{364}{365} \cdot \frac{363}{365} \cdot \dots \cdot \frac{341}{365} = \frac{364 \cdot 363 \cdot \dots \cdot 341}{365^{24}}$$

- O que dá

```
pn <- prod((364:341)/365)
pn
```

```
## [1] 0,4313003
```

- Então, $P(I)$ é

```
1 - pn
```

```
## [1] 0,5686997
```

- Surpreso? Com 25 pessoas na sala, é mais provável haver do que não haver coincidência de aniversários!

7.11.2

Simulação

- Vamos simular milhares de salas com 25 pessoas satisfazendo as **premissas** e ver em quantas delas há coincidência de aniversários. Examine o código abaixo:

```
nsalas <- 1e4
npessoas <- 25

coincidencia <- function(sala) {

  # Se a quantidade de valores únicos for diferente
  # da quantidade total de valores, então há repetição
  !(length(unique(sala)) == length(sala))

}

gerar_e_testar <- function(npessoas) {

  # Escolhemos, ao acaso, npessoas números entre 1 e 365,
  # com reposição
  sala <- sample(1:365, npessoas, replace = TRUE)

  # Testamos se há alguma coincidência de aniversários
  coincidencia(sala)

}

simular <- function(npessoas, nsalas) {

  resultados <- replicate(nsalas, gerar_e_testar(npessoas))

  # Como resultados é um vetor booleano, tirar a média
  # vai dar a proporção de resultados verdadeiros,
  # que é a probabilidade.
  mean(resultados)

}

simular(npessoas, nsalas)
```

```
## [1] 0,575
```

7.11.3

Para diferentes valores de $n \in \{2, 3, \dots, 50\}$

Soluções teóricas

- Vamos calcular as probabilidades de coincidência para diferentes quantidades n de pessoas na sala e fazer um gráfico:

```
npessoas <- 2:50

p <- function(n) {

  # Fórmula geral, para n pessoas
  1 - prod((364:(366 - n))/365)

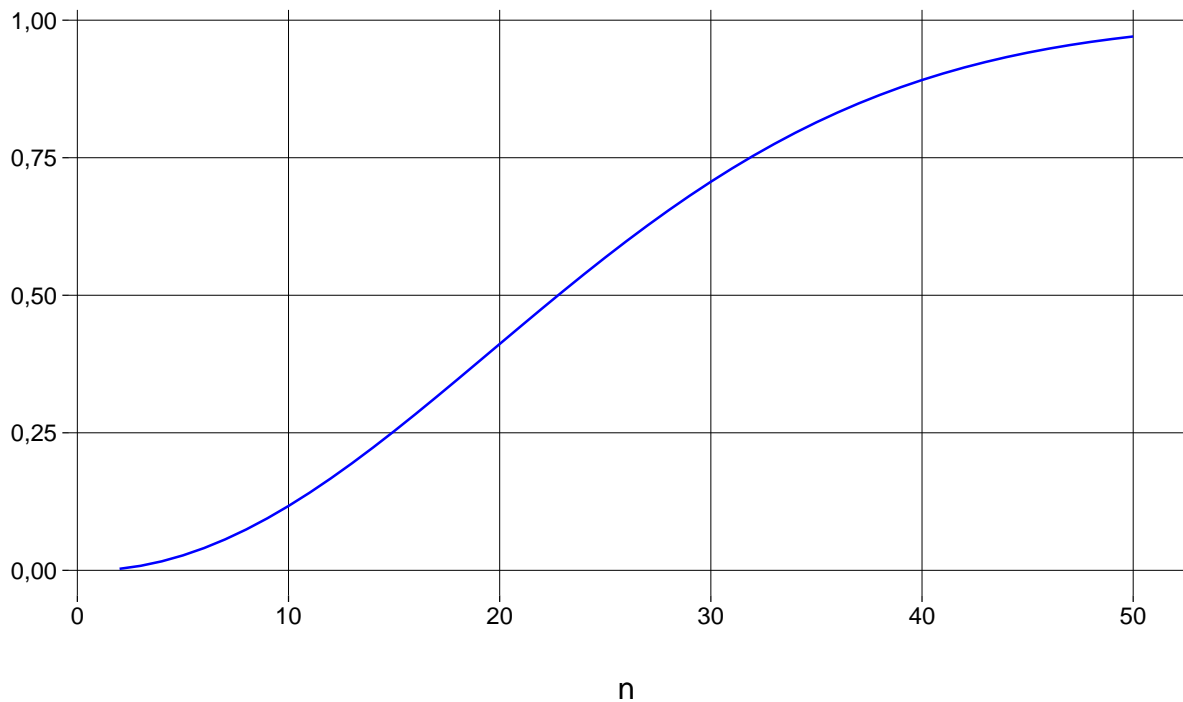
}

probs <- sapply(npessoas, p)

grafico <- probs %>%
  as_tibble() %>%
  ggplot(aes(x = npessoas, y = value)) +
    geom_line(color = 'blue') +
    labs(
      title = 'Probabilidades de coincidência com n pessoas',
      y = NULL,
      x = 'n'
    )

grafico
```

Probabilidades de coincidência com n pessoas



- Este problema é tão usado em cursos de probabilidade que o R oferece as funções `pbirthday` e `qbirthday`.
- Leia a ajuda de `pbirthday` e recrie o gráfico acima usando esta função.
- Leia a ajuda de `qbirthday` e responda:
 - Quantas pessoas são necessárias para que a probabilidade de uma ou mais coincidências seja de pelo menos 50%?
 - Quantas pessoas são necessárias para que a probabilidade de uma ou mais coincidências seja de pelo menos 90%?
 - Quantas pessoas são necessárias para que haja uma probabilidade de pelo menos 50% de que 5 ou mais pessoas façam aniversário no mesmo dia?

Simulação

```
nsalas <- 1e3
npessoas <- 2:50

probs_sim <- sapply(npessoas, simular, nsalas)

grafico +
  geom_line(
    data = as_tibble(probs_sim),
```

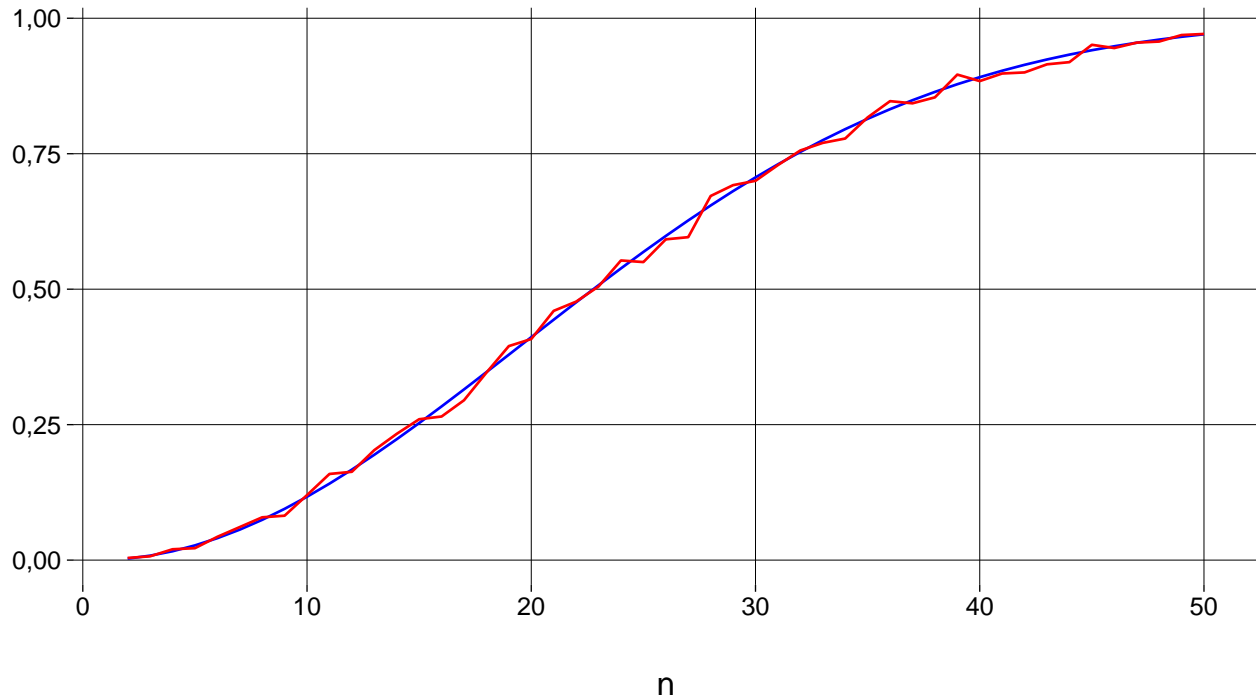
```

mapping = aes(y = value),
color = 'red'
) +
labs(
  subtitle = '(teóricas em azul, simulações em vermelho)'
)

```

Probabilidades de coincidência com n pessoas

(teóricas em azul, simulações em vermelho)



7.11.4

Premissas mais realistas

- **Vamos considerar anos bissextos.** O total de dias muda para 366, mas um dos dias (29 de fevereiro) tem $1/4$ da probabilidade de um dia normal de ser o aniversário de alguém.
- **Além disso,** vamos supor que haja 165 dias em que a probabilidade de alguém nascer é 25% maior do que nos 200 dias normais.
- A solução teórica é bem mais complexa do que no caso uniforme!
- Vamos fazer apenas a simulação.
- Preste atenção no vetor `pesos`, que representam as probabilidades de dias diferentes:
 - 200 dias normais têm peso 4;
 - 165 dias mais prováveis têm peso 5;

- 1 dia (29 de fevereiro) tem peso 1.
- Estes pesos não são probabilidades, porque a soma deles não é 1.
- A função `sample` **normaliza** automaticamente estes pesos.
- Normalizar significa dividir todos os valores pela mesma constante, de forma que a soma seja 1.

```

nsalas <- 1e3
npessoas <- 2:50

pesos <- c(
  rep(4, 200),    # dias normais
  rep(5, 165),    # dias mais prováveis
  1               # 29 de fevereiro
)

gerar_e_testar <- function(npessoas, pesos) {

  sala <- sample(1:366, npessoas, replace = TRUE, prob = pesos)
  coincidencia(sala)

}

simular <- function(npessoas, nsalas, pesos) {

  resultados <- replicate(nsalas, gerar_e_testar(npessoas, pesos))
  mean(resultados)

}

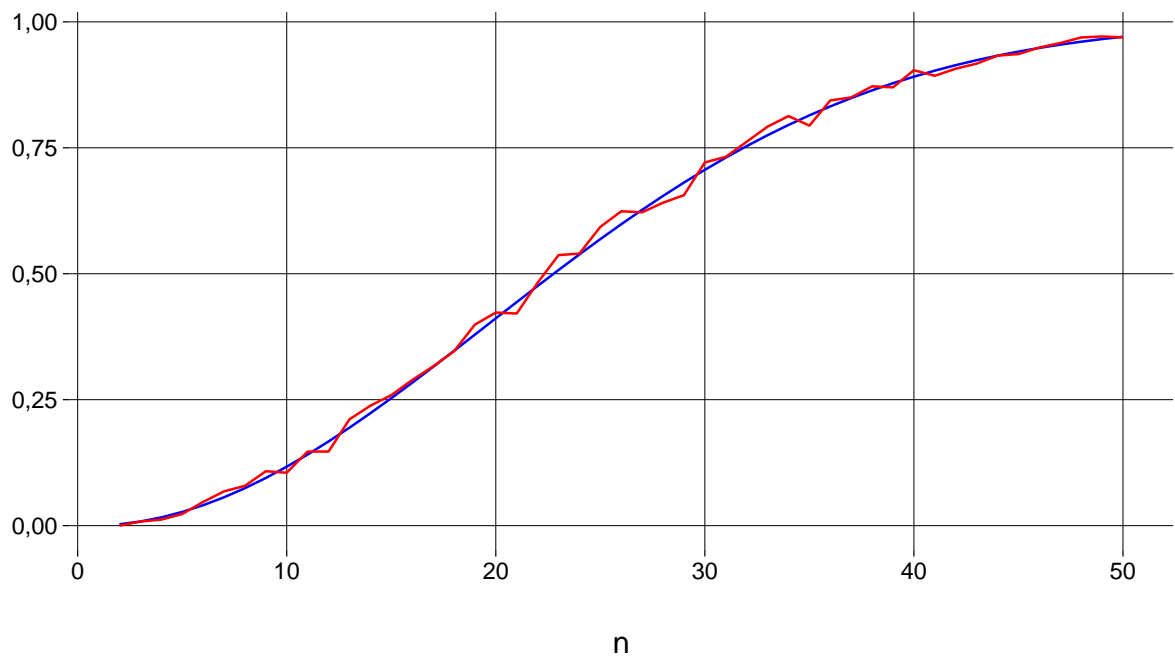
novas_probs <- sapply(npessoas, simular, nsalas, pesos)

grafico +
  geom_line(
    data = as_tibble(novas_probs),
    mapping = aes(y = value),
    color = 'red'
  ) +
  labs(
    subtitle = paste(
      '(teóricas com premissas originais: azul;',
      'simulações com novas premissas: vermelho)'
    )
  )

```

Probabilidades de coincidência com n pessoas

(teóricas com premissas originais: azul; simulações com novas premissas: vermelho)



- As novas premissas não mudaram muita coisa.
- Escreva a versão normalizada do vetor `pesos`.

7.12

Exercícios

7.12.1

Semanas com mais nascimentos

- Imagine que 50% dos nascimentos de um ano aconteçam em um período de 15 semanas, e o restante dos nascimentos seja distribuído de maneira uniforme no restante do ano. Ignore anos bissextos.
- Faça simulações como na seção anterior ($2 \leq n \leq 50$) e construa o gráfico comparando com as probabilidades teóricas (com as premissas originais).
- Interprete o resultado.

7.12.2

Pôquer

- Uma mão de pôquer consiste de 5 cartas retiradas ao acaso de um baralho de 32 cartas (4 naipes, cada um com cartas 7, 8, 9, 10, J, Q, K, A).
- Calcule as seguintes probabilidades teoricamente e através de simulações.
 1. Qual a probabilidade de obter uma mão sem ases?
 2. Qual a probabilidade de obter 4 ases?
 3. Qual a probabilidade de obter uma sequência (7 a J, 8 a Q, 9 a K, ou 10 a A) de naipes quaisquer?
 4. Qual a probabilidade de obter uma sequência (7 a J, 8 a Q, 9 a K, ou 10 a A) do mesmo naipe?

7.12.3

Dados

- Calcule as seguintes probabilidades teoricamente e através de simulações.
 1. Você lança um dado não-viciado 6 vezes. Qual a probabilidade de que saiam os 6 números?
 2. Idem, se você lançar o dado 10 vezes.

7.13

Vídeo 2

<https://youtu.be/NVP-MwtGp0Q>

7.14

Probabilidade condicional

- Em um mesmo experimento, saber que um evento B aconteceu pode dar informação sobre um outro evento A .
- Por exemplo, ao lançar um dado, a probabilidade de A — conseguir um 6 — é de $1/6$.
- Se formos informados que o evento B — o lançamento deu um número maior que 3 — ocorreu, então a probabilidade de ter conseguido um 6 passa para $1/3$.
- Escrevemos

$$P(A) = 1/6$$

e

$$P(A | B) = 1/3$$

- $P(A | B)$ é a probabilidade de A ocorrer, dado que B ocorreu.
- É uma probabilidade condicional. Estamos condicionando sobre B .

7.14.1

Exemplo: Titanic

- A seguinte tabela mostra as quantidades de pessoas no Titanic, categorizadas como sobreviventes ou não, e divididas pela classe:

##	Classe					
##	Sobreviveu	1	2	3	Tripulação	Total
##	Não	122	167	528	673	1490
##	Sim	203	118	178	212	711
##	Total	325	285	706	885	2201

Probabilidade de ser tripulante

- Escolha uma das 2.201 pessoas ao acaso.
- Qual é a probabilidade de a pessoa escolhida ser um tripulante?
- Esta é uma probabilidade não-condicional: basta dividir o total de tripulantes pelo total de pessoas:

$$P(\text{tripulante}) = \frac{885}{2.201}$$

- A tabela está na variável `tit_tab`. Em R, podemos indexar uma tabela pelos nomes. O primeiro índice corresponde à linha, o segundo à coluna:

```
ptrip <-
  tit_tab['Total', 'Tripulação'] / tit_tab['Total', 'Total']

ptrip
```

```
## [1] 0,40209
```

Probabilidade de não ser tripulante

- Escolha uma das 2.201 pessoas **ao acaso**.
- Qual é a probabilidade de a pessoa escolhida **não ser um tripulante**?
- Esta é uma probabilidade **não-condicional**: basta dividir o total de não-tripulantes pelo total de pessoas:

$$P(\text{não-tripulante}) = \frac{325 + 285 + 706}{2.201}$$

- Em R, podemos selecionar várias células de uma tabela; basta usar um vetor como índice:

```
ntrip <- sum(tit_tab['Total', c('1', '2', '3')])  
  
ntrip / tit_tab['Total', 'Total']
```

```
## [1] 0,59791
```

- Mas nem era preciso fazer este cálculo. Basta perceber que “ser tripulante” e “ser não-tripulante” são **eventos complementares**. Daí,

$$P(\text{não-tripulante}) = 1 - P(\text{tripulante})$$

```
1 - ptrip
```

```
## [1] 0,59791
```

Probabilidade de sobreviver

- Escolha uma das 2.201 pessoas **ao acaso**.
- Qual é a probabilidade de a pessoa escolhida **ter sobrevivido**?
- Esta é uma probabilidade **não-condicional**: basta dividir o total de sobreviventes pelo total de pessoas:

$$P(\text{sobrevivente}) = \frac{711}{2.201}$$

```
tit_tab['Sim', 'Total'] / tit_tab['Total', 'Total']
```

```
## [1] 0,323035
```

Probabilidade de ser de primeira classe

- Escolha uma das 2.201 pessoas **ao acaso**.
- Qual é a probabilidade de a pessoa escolhida **ser da primeira classe**?
- Esta é uma probabilidade **não-condicional**: basta dividir o total de passageiros da primeira classe pelo total de pessoas:

$$P(1^{\text{a}} \text{ classe}) = \frac{325}{2.201}$$

```
tit_tab['Total', '1'] / tit_tab['Total', 'Total']
```

```
## [1] 0,1476602
```

Probabilidade de sobreviver E ser de primeira classe

- Escolha uma das 2.201 pessoas **ao acaso**.
- Qual é a probabilidade de a pessoa escolhida **ter sobrevivido e ser da primeira classe**?
- Isto é uma **probabilidade conjunta** — a probabilidade de dois eventos terem ocorrido ao mesmo tempo. **Ainda não é uma probabilidade condicional**.
- Queremos saber a proporção de pessoas, do total de pessoas a bordo, que eram de primeira classe e sobreviveram.

$$P(\text{sobrevivente da } 1^{\text{a}} \text{ classe}) = \frac{203}{2.201}$$

```
tit_tab['Sim', '1'] / tit_tab['Total', 'Total']
```

```
## [1] 0,0922308
```

Probabilidade de uma pessoa da primeira classe sobreviver

- Escolha uma das 2.201 pessoas **ao acaso**.
- Qual é a probabilidade de a pessoa escolhida **ter sobrevivido, dado que a pessoa estava na primeira classe?**
- Isto é uma **probabilidade condicional**, escrita como

$$P(\text{sobrevivente} \mid 1^{\text{a}} \text{ classe})$$

- Cuidado, agora.
- Já sabemos que a pessoa é da primeira classe. Logo, **restringimos o universo a estas 325 pessoas**. O denominador vai ser o total de passageiros da primeira classe:

$$P(\text{sobrevivente} \mid 1^{\text{a}} \text{ classe}) = \frac{203}{325}$$

```
tit_tab['Sim', '1'] / tit_tab['Total', '1']
```

```
## [1] 0,6246154
```

- Perceba que

$$P(\text{sobreviveu} \mid 1^{\text{a}} \text{ classe})$$

é o mesmo que

$$\frac{P(\text{sobreviveu} \cap 1^{\text{a}} \text{ classe})}{P(1^{\text{a}} \text{ classe})}$$

Probabilidade de um sobrevivente ser da primeira classe

- Escolha uma das 2.201 pessoas **ao acaso**.
- Qual é a probabilidade de a pessoa escolhida **ser da primeira classe, dado que ela sobreviveu?**
- Isto é outra **probabilidade condicional**, escrita como

$$P(1^{\text{a}} \text{ classe} \mid \text{sobreviveu})$$

- **Não é a mesma probabilidade** que $P(\text{sobreviveu} \mid 1^{\text{a}} \text{ classe})$.

- Em português:
 - $P(\text{sobreviveu} \mid 1^{\text{a}} \text{ classe})$ é a probabilidade de
 - * A pessoa sobreviver, dado que era da primeira classe;
 - * Equivalentemente: alguém da primeira classe sobreviver.
 - $P(1^{\text{a}} \text{ classe} \mid \text{sobreviveu})$ é a probabilidade de
 - * A pessoa ter sido da primeira classe, dado que sobreviveu;
 - * Equivalentemente: alguém que sobreviveu ter sido da primeira classe.
 - Releia até entender.
- Agora, restringimos o universo às pessoas que sobreviveram. Dentre estas, quantas são da primeira classe?

$$P(1^{\text{a}} \text{ classe} \mid \text{sobreviveu}) = \frac{203}{711}$$

```
tit_tab['Sim', '1'] / tit_tab['Sim', 'Total']
```

```
## [1] 0,2855134
```

Este é um exemplo de que $P(A \mid B)$ pode ser diferente de $P(B \mid A)$.

7.14.2

Definição de probabilidade condicional

- Como vimos nos exemplos, para calcular $P(A \mid B)$, restringimos o universo aos elementos onde B acontece, e, deste universo, verificamos quantos elementos também correspondem a A acontecer — isto é, elementos onde $A \cap B$ acontece.
- Em termos de frequência relativa:

$$\frac{\text{ocorrências de } A \cap B}{\text{ocorrências de } B}$$

- Em termos de probabilidade, a definição é

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

7.14.3

Exercícios

No Titanic,

1. Qual a probabilidade de um tripulante sobreviver?
2. Qual a probabilidade de um sobrevivente ser tripulante?
3. Qual a probabilidade de um não-tripulante sobreviver?
4. Qual a probabilidade de um sobrevivente não ser tripulante?
5. Compare as probabilidades condicionais de uma pessoa sobreviver dado que
 - a. Ela estava na 1ª classe. (Já calculada no exemplo acima: 0,62.)
 - b. Ela estava na 2ª classe.
 - c. Ela estava na 3ª classe.
 - d. Ela era da tripulação.

Que conclusões você tira?

7.15

Probabilidade conjunta

- Imagine que queremos calcular a probabilidade de dois eventos A e B acontecerem **ao mesmo tempo**.
- Ou seja, queremos descobrir **a probabilidade conjunta $P(A \cap B)$** .
- Muitas vezes, é difícil calcular esta probabilidade diretamente.
- A fórmula para calcular $P(A | B)$ nos dá uma maneira de calcular $P(A \cap B)$:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \iff P(A \cap B) = P(A | B) \cdot P(B)$$

- Ou, invertendo A e B ,

$$P(B | A) = \frac{P(A \cap B)}{P(A)} \iff P(A \cap B) = P(B | A) \cdot P(A)$$

- Em palavras:
 - A **probabilidade de A e B acontecerem juntos** é a **probabilidade de A dado que B aconteceu**, multiplicada pela **probabilidade de B** .
 - Ou, invertendo A e B , a **probabilidade de A e B acontecerem juntos** é a **probabilidade de B dado que A aconteceu**, multiplicada pela **probabilidade de A** .

7.16

Independência

- Mais acima, vimos que, para dois eventos independentes A e B , a probabilidade conjunta $P(A \cap B)$ é igual a $P(A) \cdot P(B)$.
- Olhando para as fórmulas acima para a probabilidade conjunta, se A e B forem independentes, então

$$\begin{aligned} P(A \cap B) &= P(A \mid B) \cdot P(B) \\ &= P(A) \cdot P(B) \end{aligned}$$

o que nos diz que

$$P(A \mid B) = P(A)$$

- Da mesma forma,

$$P(B \mid A) = P(B)$$

- Em palavras: saber que um dos eventos ocorreu não altera a probabilidade do outro evento.
- Qualquer uma das 3 igualdades pode ser tomada como a definição formal de eventos independentes.

7.16.1

Exemplos

Estar na primeira classe e sobreviver são independentes?

- A probabilidade de sobreviver, dado que a pessoa estava na 1ª classe, é

$$P(\text{sobreviver} \mid 1^{\text{a}} \text{ classe}) = \frac{203}{325}$$

```
tit_tab['Sim', '1'] / tit_tab['Total', '1']
```

```
## [1] 0,6246154
```

- Mas a probabilidade (incondicional) de sobreviver é

$$P(\text{sobreviver}) = \frac{711}{2.201}$$

```
tit_tab['Sim', 'Total'] / tit_tab['Total', 'Total']
```

```
## [1] 0,323035
```

- Como as probabilidades são diferentes, os eventos **não são independentes**.
- Verifique se $P(\text{sobreviver} \cap 1^{\text{a}} \text{ classe}) = P(\text{sobreviver}) \cdot P(1^{\text{a}} \text{ classe})$.

Faltas e turno de trabalho

- Numa empresa:
 - 75 funcionários trabalham no turno diurno, com um número de faltas de 3 por semana.
 - 25 funcionários trabalham no turno noturno, com um número de faltas de 1 por semana.
 - Faltar é independente do turno de trabalho?
- Vamos construir uma tabela:

```
faltas <- array(
  c(3, 1, 72, 24),
  dim = c(2, 2)
) %>%
  addmargins()

dimnames(faltas) = list(
  'Turno' = c('Diurno', 'Noturno', 'Total'),
  'Presença' = c('Faltou', 'Presente', 'Total')
)

faltas
```

```
##          Presença
## Turno      Faltou Presente Total
## Diurno       3         72    75
## Noturno       1         24    25
## Total        4         96   100
```

- A probabilidade (incondicional) de faltar é

$$P(\text{Faltou}) = \frac{4}{100}$$

```
faltas['Total', 'Faltou'] / faltas['Total', 'Total']
```

```
## [1] 0,04
```

- A probabilidade de faltar no turno diurno é

$$P(\text{Faltou} \mid \text{Diurno}) = \frac{3}{75}$$

```
faltas['Diurno', 'Faltou'] / faltas['Diurno', 'Total']
```

```
## [1] 0,04
```

- Como as probabilidades são iguais, os eventos **são independentes**.
- Verifique que $P(\text{Faltou}) = P(\text{Faltou} \mid \text{Noturno})$.

7.17

Probabilidade total

7.17.1

Exemplo

- Dentre 80 homens, 30 têm olhos azuis.
- Dentre 50 mulheres, 20 têm olhos azuis.
- Neste grupo de pessoas, qual a **probabilidade de ter olhos azuis**?
- Homens e mulheres formam uma **partição** deste grupo — i.e., cada pessoa **só pode** ser homem ou mulher (não ambos) e cada pessoa **precisa** ser homem ou mulher.
- O evento “ter olhos azuis” se subdivide em dois casos **mutuamente exclusivos**:
 1. Ter olhos azuis e ser homem;
 2. Ter olhos azuis e ser mulher.
- Vamos chamar os eventos de
 - A = ter olhos azuis
 - H = ser homem
 - M = ser mulher
- Calculamos a probabilidade $P(A)$ somando as probabilidades dos dois casos:

$$P(A) = P(A \cap H) + P(A \cap M)$$

- Então,

$$P(A) = \frac{30}{130} + \frac{20}{130} = \frac{50}{130} \approx 0,39$$

- Ou, **antes de somar**, podemos transformar as probabilidades conjuntas em produtos de uma probabilidade condicional por uma probabilidade não-condicional:

$$\begin{aligned} P(A) &= P(A \cap H) + P(A \cap M) \\ &= P(A | H)P(H) + P(A | M)P(M) \end{aligned}$$

- O que nos dá o mesmo resultado:

$$\begin{aligned} P(A) &= \frac{30}{80} \cdot \frac{80}{130} + \frac{20}{50} \cdot \frac{50}{130} \\ &= \frac{50}{130} \\ &\approx 0,39 \end{aligned}$$

7.18

Teorema de Bayes

7.18.1

Exemplo

- De **todos** os *emails*, 60% são *spam*:

$$P(\text{spam}) = 0,6$$

- De todos os *emails* **que são spam**, 90% contêm a palavra “compre”:

$$P(\text{compre} | \text{spam}) = 0,9$$

- De **todos** os *emails* (*spam* ou não), 70% contêm a palavra “compre”:

$$P(\text{compre}) = 0,7$$

- Você acaba de receber um *email*. **Antes de você abri-lo, qual a probabilidade de o *email* ser *spam*?**
- Bem, **na ausência de informação adicional**, $P(\text{spam}) = 0,6$.
- Você abre o *email*. **Ele contém a palavra “compre”.**

- Agora, qual a probabilidade de ser *spam*?

$$P(\text{spam} \mid \text{compre}) = ?$$

- Lembre-se de que

$$P(\text{compre} \cap \text{spam}) = P(\text{spam} \mid \text{compre}) \cdot P(\text{compre})$$

- Mas também

$$P(\text{compre} \cap \text{spam}) = P(\text{compre} \mid \text{spam}) \cdot P(\text{spam})$$

- As duas expressões são iguais:

$$P(\text{spam} \mid \text{compre}) \cdot P(\text{compre}) = P(\text{compre} \mid \text{spam}) \cdot P(\text{spam})$$

- Isolando o termo que queremos descobrir:

$$P(\text{spam} \mid \text{compre}) = \frac{P(\text{compre} \mid \text{spam}) \cdot P(\text{spam})}{P(\text{compre})}$$

- Substituindo os valores:

$$P(\text{spam} \mid \text{compre}) = \frac{0,9 \cdot 0,6}{0,7} \approx 0,77$$

- Isto é **inferência bayesiana**:

1. Começamos com uma probabilidade não-condicional: **a priori**, $P(\text{spam}) = 0,6$;
2. Obtivemos **nova informação**: o *email* contém “compre”;
3. Usamos esta informação para calcular uma probabilidade condicional, **a posteriori**:

$$P(\text{spam} \mid \text{compre}) = \frac{P(\text{compre} \mid \text{spam}) \cdot P(\text{spam})}{P(\text{compre})}$$

- Perceba que, para isso, precisamos da probabilidade não-condicional $P(\text{compre})$ (no denominador).

7.18.2

No geral

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

- E se você não souber $P(B)$?
- Use o teorema da probabilidade total:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n)$$

onde os A_i formam uma partição da população.

- Isto equivale a

$$P(B) = P(B | A_1)P(A_1) + P(B | A_2)P(A_2) + \dots + P(B | A_n)P(A_n)$$

7.18.3

Outro exemplo

- Uma **doença rara** afeta 5 pessoas em 100.000.
- O exame que detecta a doença tem precisão de 99,9%; i.e., **quando uma pessoa está doente, o exame dá positivo 99,9% das vezes.**
- **Quando uma pessoa não está doente, o exame dá positivo 1% das vezes.** Este caso é um falso positivo.
- Você faz o exame, e **o resultado é positivo.**
- **Dado este resultado, qual a probabilidade de você ter a doença?**
- Vamos nomear os eventos:
 - D = você está doente;
 - ND = você **não** está doente;
 - $+$ = o exame deu positivo;
 - $-$ = o exame deu negativo.
- Vamos usar Bayes:

$$P(D | +) = \frac{P(+ | D)P(D)}{P(+)}$$

- $P(+ | D) = 0,999$, pelo enunciado.
- $P(D) = 0,00005$, pelo enunciado.

- Daí, $P(ND) = 1 - P(D) = 0,99995$.
- Não temos $P(+)$, mas podemos achar usando o **teorema da probabilidade total**, lembrando que o enunciado diz que $P(+ | ND) = 0,01$:

$$\begin{aligned}
 P(+) &= P(+ \cap D) + P(+ \cap ND) \\
 &= P(+ | D)P(D) + P(+ | ND)P(ND) \\
 &= 0,999 \cdot 0,00005 + 0,01 \cdot 0,99995 \\
 &= 0,01004945
 \end{aligned}$$

- Inserindo os valores no teorema de Bayes:

$$\begin{aligned}
 P(D | +) &= \frac{P(+ | D)P(D)}{P(+)} \\
 &= \frac{0,999 \cdot 0,00005}{0,01004945} \\
 &= 0,00497
 \end{aligned}$$

- A probabilidade de estar doente é **menos do que 0,5%**!
- Você provavelmente esperava uma probabilidade maior.
- Qual das 3 probabilidades usadas no cálculo fez o resultado ser tão pequeno?
 - $P(+ | D)$?
 - $P(D)$?
 - $P(+)$?

Referências

Oliveira Morgado, Augusto César de, João Bosco Pitombeira de Carvalho, Paulo Cezar Pinto Carvalho, e Pedro Jesus Fernandez. 2004. *Análise combinatória e probabilidade*. Rio de Janeiro: Impa / Vitae.