Lição de Casa 1

James R. Hunter, PhD.

18 de setembro de 2020

Nesta lição da casa, vamos trabalhar com algumas problemas verdadeiras e alguns simplificados. As respostas precisam ser submetidas antes de **2 de outubro** por email: jameshunterbr@gmail.com.

Problema 1

O arquivo `trplasma.csv` (no GitHub) mostra todas as mutações que uma amostra dos pacientes com HIV em falha virológica tiveram no gene transcriptase reversa do vírus. Um "1" na célula da planilha indica que a mutação estava presente e "0" demonstra ausência da mutação naquele momento.

A primeira coluna é o número de código do paciente e todos as outras colunas representam as mutações. Os nomes das colunas podem ser interpretados como:

- "tr" para *transcriptase reversa*,
- o número seguinte como a posição do aminoácido (codon) com a mutação
- o código para semana de exame de sangue ("bl" = *baseline* ou "_12" = 12 semanas).

Primeira tarefa: importar o arquivo usando `readr::read_csv()` para um tibble chamada `trplas`.

Segunda tarefa: Transforme os dados para um conjunto *tidy*, usando as funções de `tidyr`.

Dica: planeje a transformação. Quais são as variáveis? Como vai dividir o código da mutação? Desenhe no papel como vai aparecer o tibble.

Solution

```
library(tidyverse, quietly = TRUE)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ ti
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Primeira tarefa
trplas <- readr::read_csv(here::here("trplasma.csv"), col_names = TRUE,
                          col_types = "cfffffffffffffffffffffffffffffffffffffffffffffff")
# Segunda tarefa
trplas_tidy <- trplas %>%
  tidyr::pivot_longer(., cols = tr41bl:tr219_12, names_to = "mutacao", values_to = "value")
tibble::glimpse(trplas_tidy)
```

```
## Rows: 1,748
## Columns: 3
## $ n       <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",...
## $ mutacao <chr> "tr41bl", "tr44bl", "tr65bl", "tr67bl", "tr69bl", "tr70bl",...
## $ value   <fct> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...
```

Problema 2

Carregue na memoria o arquivo "pac_demo.xlsx" (no GitHub) como `pac_data`, e responda às perguntas seguintes.

a. Existe uma diferença entre as médias de `copias_cv` entre pacientes de São Paulo e de Rio Grande de Sul? Quanto?

b. Quando falamos de carga viral de HIV, normalmente falamos de uma tranformação das cópias do vírus em log10. Usando a função `log10()`, crie uma nova variável da valor logaritmico de carga viral, `log_cv` e salvar ele de volta para `pac_data`.

c. O que é a idade mediana das pacientes com baixa contagem das células CD4+ (`contagem_cd4`), ou seja, um valor em baixo de 200?

d. Este valor é maior ou menor da idade mediana das pessoas com contagem de CD4+ maior de 200? Porque você acha que esta diferença existe?

```
pac_data <- readxl::read_excel(here::here("pac_demo.xlsx"))

# Parte a
x <- pac_data %>%
  filter(ufnasc %in% c("RS", "SP")) %>%
  group_by(ufnasc) %>%
  summarise(mean_copias = mean(copias_cv)) %>%
  ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
(difference <- abs(x[1,2] - x[2,2]))
```

```
##   mean_copias
## 1    104747.1
```

```
# Parte b

pac_data$log_cv <- log10(pac_data$copias_cv)

# Parte c

low_cd <- as.numeric(pac_data %>%
  filter(contagem_cd4 < 200) %>%
  summarise(med = median(idade)))

# Parte d

hi_cd <- as.numeric(pac_data %>%
  filter(contagem_cd4 >= 200) %>%
  summarise(med = median(idade)))

paste("A idade das pessoas com CD4+ < 200 é", abs(low_cd - hi_cd), "anos menos")
```

```
## [1] "A idade das pessoas com CD4+ < 200 é 4.5 anos menos"
```

Problema 3

No pacote `nycflights13`, o pior linha aérea em demoras de chegada nos destinos foi Frontier Airlines. Agora vamos considerar o destino que teve o pior desempenho em termos de demora. O diagrama da estrutura de `nycflights13` fica nos slides de Aula 3.

Qual é o tempo médio de demora na chegada para todos os destinos que os vôos de Nova York e o que é o nome do pior?

```
library(nycflights13)

flights <- nycflights13::flights %>%
  select(arr_delay, dest)
airports <- nycflights13::airports %>%
  select(faa, name)

# tempo médio de demora para todos os vôos

mean(flights$arr_delay, na.rm = TRUE)
```

```
## [1] 6.895377
```

```
# aeroporto de chegada pior

worst <- flights %>%
  filter(!is.na(arr_delay)) %>%
  left_join(airports, by = c("dest" = "faa")) %>%
  group_by(name) %>%
  summarise(dem_med = mean(arr_delay)) %>%
  ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
worst %>%
  filter(dem_med == max(dem_med))
```

```
## # A tibble: 1 x 2
##   name                  dem_med
##   <chr>                   <dbl>
## 1 Columbia Metropolitan    41.8
```