Laboratório 01

Oscar J. O. Ayala

Introdução

Um conjunto de dados no formato *tidy* beneficia o analista de dados por permitir a manipulação dos mesmos de uma maneira unificada. De modo similar, métodos estatísticos são habitualmente implementados para receber dados neste formato. Desta maneira, a importação e tratamento de dados visando o referido formato reduzirá a criação de bancos de dados temporários, evitando problemas difíceis de diagnosticar.

Os conjuntos de dados apresentados correspondem ao número de casos de tuberculose observados em alguns países, juntamente com seus tamanhos populacionais.

Manipulação de Dados no Formato Tidy

- 1. Carregue o pacote `tidyverse`

```
library(tidyverse)
```

- 2. Apresente os bancos de dados `table1` , `table2` , `table3` , `table4a` e `table4b` , distribuídos juntamente com o pacote `tidyverse` . Para cada banco de dados, descreva textualmente se ele está no formato tidy e justifique cada uma de suas respostas.

A `Table 1` está em formato *tidy*, cada coluna uma variável, cada linha um registro e cada célula uma única entrada.

```
tidyr::table1 %>%
  knitr::kable(caption = "Table 1. Tidy format")
```

Table 1: Table 1. Tidy format

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

A `Table 2`, não esta em formato *tidy*, tem mais de uma variável no campo *type*.

```
tidyr::table2 %>%
  knitr::kable(caption = "Table 2. Not tidy format")
```

Table 2: Table 2. Not tidy format

country	year	type	count
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

A `Table 3`, não esta em formato *tidy*, tem mais de uma variável no campo *type*.

```
tidyr::table3 %>%
  knitr::kable(caption = "Table 3. Not tidy format")
```

Table 3: Table 3. Not tidy format

country	year	rate
Afghanistan	1999	745/19987071
Afghanistan	2000	2666/20595360
Brazil	1999	37737/172006362
Brazil	2000	80488/174504898
China	1999	212258/1272915272
China	2000	213766/1280428583

As `Table 4a` e `Table 4b`, não estão em formato *tidy*, os campos `1999` e `2000` não são variáveis.

```
tidyr::table4a %>%
  knitr::kable(caption = "Table 4a. Not tidy format")
```

Table 4: Table 4a. Not tidy format

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

```
tidyr::table4b %>%
  knitr::kable(caption = "Table 4b. Not tidy format")
```

Table 5: Table 4b. Not tidy format

country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

- 3. Utilizando comandos do pacote `dplyr` , determine a taxa de ocorrência de tuberculose para cada `10.000` pessoas. Armazene o resultado em um objeto chamado `taxas`.

```
taxas <- with(tidyr::table1, cases / population * 10000)
taxas
```

```
## [1] 0.372741 1.294466 2.193930 4.612363 1.667495 1.669488
```

- 4. Apresente, utilizando comandos do pacote `dplyr` , o número de casos de tuberculose por ano.

```
tidyr::table1 %>% dplyr::group_by(year) %>%
  dplyr::summarise(total_cases = sum(cases))
```

```
## # A tibble: 2 x 2
##    year total_cases
##   <int>       <int>
## 1  1999      250740
## 2  2000      296920
```

- 5. Apresente, utilizando comandos do pacote `dplyr`, o número de casos de tuberculose identificados em cada país.

```
tidyr::table1 %>%
  dplyr::group_by(country) %>%
  dplyr::summarise(total_cases = sum(cases))
```

```
## # A tibble: 3 x 2
##   country     total_cases
##   <chr>             <int>
## 1 Afghanistan        3411
## 2 Brazil           118225
## 3 China            426024
```

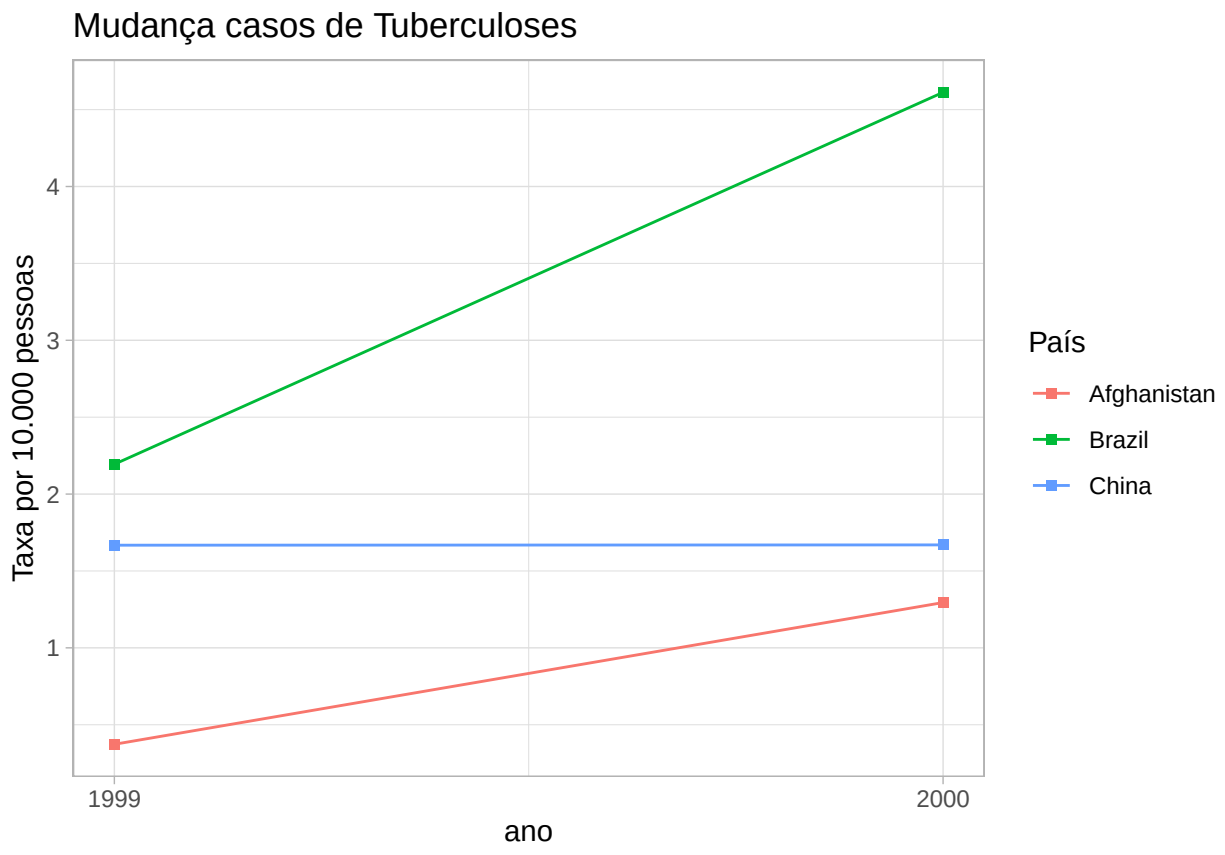
- 6. Utilizando comandos do pacote `dplyr` , apresente uma tabela que descreva a mudança no número de casos, em cada país, ao longo dos anos de `1999` e `2000`.

```
tidyr::table1 %>% dplyr::select(-population)
```

```
## # A tibble: 6 x 3
##   country      year  cases
##   <chr>       <int>  <int>
## 1 Afghanistan  1999    745
## 2 Afghanistan  2000   2666
## 3 Brazil       1999  37737
## 4 Brazil       2000  80488
## 5 China        1999 212258
## 6 China        2000 213766
```

- 7. Apresente um gráfico de linhas, preparado via `ggplot2` , apresentando a mudança na taxa de casos (por `10.000` habitantes) estratificado por país.

```
tidyr::table1 %>%
  dplyr::mutate(taxas = cases / population * 10000)  %>%
  ggplot2::ggplot(ggplot2::aes(year, taxas, group = country)) +
  ggplot2::geom_line(ggplot2::aes(color = country)) +
  ggplot2::geom_point(ggplot2::aes(color = country), shape = 15) +
  ggplot2::scale_x_continuous(breaks = c(1999, 2000)) +
  ggplot2::labs(title = "Mudança casos de Tuberculoses", x = "ano",
                y = "Taxa por 10.000 pessoas") +
  ggplot2::scale_color_discrete("País") +
  ggplot2::theme_light()
```



- 8. Calcule a taxa para as tabelas `table2` e `table4a + table4b`.

```
table2 %>% tidyr::pivot_wider(id_cols = c("country", "year"),
                              names_from = type, values_from = count) %>%
  dplyr::mutate(taxas = cases / population * 10000) %>%
  knitr::kable(caption = "Formato tidy")
```

Table 6: Formato tidy

country	year	cases	population	taxas
Afghanistan	1999	745	19987071	0.372741
Afghanistan	2000	2666	20595360	1.294466
Brazil	1999	37737	172006362	2.193931
Brazil	2000	80488	174504898	4.612363
China	1999	212258	1272915272	1.667495
China	2000	213766	1280428583	1.669488

```
library(magrittr)

table4a %<>% tidyr::pivot_longer(cols = c("1999", "2000"), names_to = "year",
                                 values_to = "cases", values_transform = as.integer)

table4b %<>% tidyr::pivot_longer(cols = c("1999", "2000"), names_to = "year",
                                 values_to = "population", values_transform = as.integer)

table_ab <- dplyr::inner_join(table4a, table4b, by = c("country", "year"))
table_ab %>% knitr::kable(caption = "Inner join")
```

Table 7: Inner join

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

- 9. Observe que a coluna rate do objeto `table3` é um texto mostrando a fração que formaria a taxa de casos de tuberculose. Transforme o objeto `table3` em um objeto com formato `tidy` separando a coluna 3 em duas outras colunas: cases e population, utilizando o comando `separate` . Utilize o argumento `convert` para transformar o resultado em um objeto numérico.

```
table3 %>%
  tidyr::separate(col = rate, into = c("cases", "population"),
                  sep = "/", convert = TRUE) %>%
  knitr::kable(caption = "Formato tidy")
```

Table 8: Formato tidy

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

A experiência de usar o *R* é única, muito boa. Fim!