Relatório de dados de Sífilis

Mikael Marin Coletto

Índice

1 Introdução

Este relatório tem como objetivo apresentar os dados de Sífilis no Brasil. Os dados foram obtidos do Sistema de Informação de Agravos de Notificação (SINAN) e referem-se ao período de 2007 a 2023.

2 Dados

Os dados foram obtidos do banco de dados do SINAN e estão disponíveis no arquivo `df_sifilis_mun_ano_fx_et_racacor_esc.RDS`. O arquivo contém as seguintes variáveis:

- Município de residência
- Ano do diagnóstico
- Mês do diagnóstico
- Faixa etária
- Raça/cor
- Escolaridade
- Número de casos de Sífilis

```
df_sifilis <- readRDS("data-raw/df_sifilis_mun_ano_fx_et_racacor_esc.RDS")
head(df_sifilis) |>
  gt::gt()
```

uf_res	mun_res	n_casos_sifg	nasc_viv	nasc_c_anom	ano_diag	faixa_etaria	raca
32	320530	1	0	0	****	25-29	Branca
33	330455	1	0	0	****	Em branco/Inválido	Parda
32	320530	1	0	0	****	25-29	Em branco/I
32	320530	1	0	0	****	30-39	Parda
32	320530	2	0	0	****	20-24	Parda
32	320332	1	0	0	****	20-24	Branca

3 Análise Exploratória

3.1 Completude de variáveis

```
variaveis <- colnames(df_sifilis)

library(dplyr)

## Definindo como NA as variáveis que possuem valores inválidos
df_sifilis_test <- df_sifilis %>%
  dplyr::mutate(across(where(is.character), ~ ifelse(.x %in% c("****", ""), NA, .x)))

df_completude <- data.frame(variavel = character(), completude = numeric())
for(i in variaveis){
  completude = round(sum(!is.na(df_sifilis_test[[i]]))/nrow(df_sifilis_test), 2)
  df_completude <- rbind(df_completude, data.frame(variavel = i, completude = completude))
}
```

```r
df_completude |>
  gt::gt() |>
  gt::cols_label(variavel = "Variável", completude = "Completude")
```

Variável	Completude
uf_res	1
mun_res	1
n_casos_sifg	1
nasc_viv	1
nasc_c_anom	1
ano_diag	1
faixa_etaria	1
raca	1
escolaridade	1

3.2 Variáveis numéricas

```r
df_sifilis_numeric_vars <- df_sifilis_test |>
  dplyr::select(where(is.numeric))

df_numeric <- data.frame(variavel = character(),
                         min = numeric(), median = numeric(), media = numeric() , max = nume
                         total = numeric())

variaveis <- colnames(df_sifilis_numeric_vars)
for(i in variaveis){
  min = min(df_sifilis_numeric_vars[[i]], na.rm = TRUE)
  median = median(df_sifilis_numeric_vars[[i]], na.rm = TRUE)
  media = round(mean(df_sifilis_numeric_vars[[i]], na.rm = TRUE), 2)
  max = max(df_sifilis_numeric_vars[[i]], na.rm = TRUE)
  sum = sum(df_sifilis_numeric_vars[[i]], na.rm = TRUE)
  df_numeric <- rbind(df_numeric, data.frame(variavel = i,
                                             min = min, median = median, media = media,
                                             max = max, total = sum))
}

df_numeric |>
```

```
  gt::gt() |>
  gt::cols_label(variavel = "Variável", min = "Mínimo", median = "Mediana", media = "Média",
```

Variável	Mínimo	Mediana	Média	Máximo	Total
n_casos_sifg	0	0	0.16	640	618921
nasc_viv	0	2	12.79	60026	48257686
nasc_c_anom	0	0	0.10	890	392640

3.3 Variáveis categóricas

```
df_sifilis_categoric_vars <- df_sifilis |>
  dplyr::select(where(is.character))

variaveis <- colnames(df_sifilis_categoric_vars)

df_categoric <- data.frame(variavel = character(),
                           n_distinct = numeric(),
                           levels = character())

for(i in variaveis){
  n_distinct = n_distinct(df_sifilis_categoric_vars[[i]], na.rm = TRUE)
  levels = paste(sort(unique(df_sifilis_categoric_vars[[i]])), collapse = ", ")
  if(length(unique(df_sifilis_categoric_vars[[i]])) > 20){
    levels = "Muitos valores distintos (> 20)"
  }
  df_categoric <- rbind(df_categoric, data.frame(variavel = i,
                                                 n_distinct = n_distinct, levels = levels))
  # print(df_categoric)
}

df_categoric |>
  gt::gt() |>
  gt::cols_label(variavel = "Variável", n_distinct = "Nº de valores distintos", levels = "Va
```

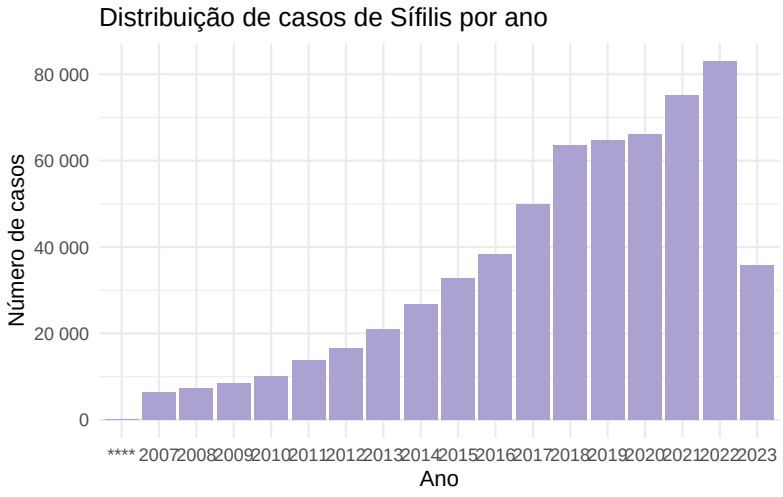
Variável	Nº de valores distintos	Valores distintos
uf_res	27	Muitos valores distintos (> 20)
mun_res	5608	Muitos valores distintos (> 20)

ano_diag	18	****, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2
faixa_etaria	8	10-14, 15-19, 20-24, 25-29, 30-39, 40-59, 60+, Em branco/Inválido
raca	6	Amarela, Branca, Em branco/Inválido, Indigena, Parda, Preta
escolaridade	12	1ª a 4ª série incompleta do EF, 4ª série completa do EF, 5ª à 8ª série in

3.3.1 Distribuição de número de casos de Sífilis

3.3.1.1 Ano de diagnóstico

```
df_sifilis_ano <- df_sifilis |>
  dplyr::group_by(ano_diag) |>
  dplyr::summarise(n_casos_sifg = sum(n_casos_sifg)) |>
  dplyr::ungroup()

df_sifilis_ano |>
  ggplot2::ggplot() +
  ggplot2::geom_col(aes(x = ano_diag, y = n_casos_sifg), fill = "#ABA2D1") +
  ggplot2::labs(title = "Distribuição de casos de Sífilis por ano",
       x = "Ano",
       y = "Número de casos") +
  ggplot2::theme_minimal() +
  ggplot2::scale_y_continuous(labels = scales::number)
```

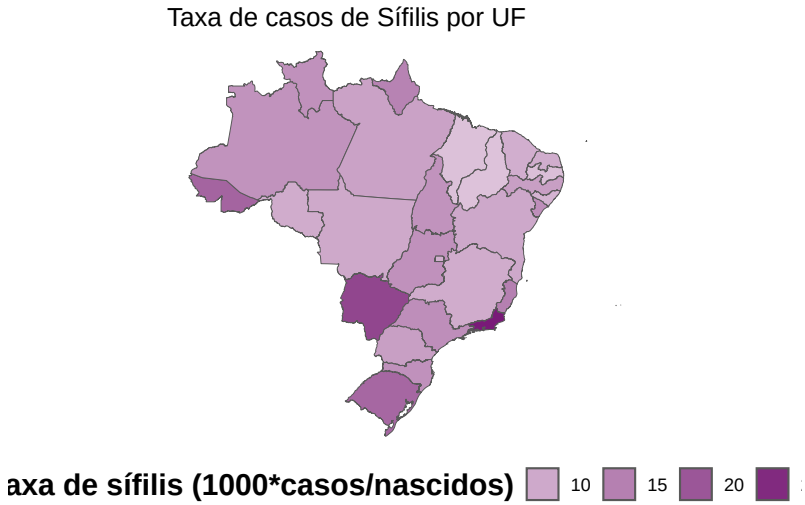


3.3.1.2 Nos estados

```r
df_sifilis_uf <- df_sifilis |>
  dplyr::group_by(uf_res) |>
  dplyr::summarise(n_casos_sifg = sum(n_casos_sifg),
                   n_nascidos = sum(nasc_viv)) |>
  dplyr::ungroup() |>
  dplyr::mutate(taxa_sifilis = n_casos_sifg / n_nascidos * 1000)

df_ufs <- data.table::fread("data-raw/ibge-ufs-pop-2022-est.csv") |>
  dplyr::mutate(uf_cod = as.character(uf_cod)) |>
  dplyr::select(uf_cod, uf_nome)

df_sifilis_uf <- dplyr::left_join(df_sifilis_uf, df_ufs, by = c("uf_res" = "uf_cod"))

uf_sf <- sf::read_sf(here::here("data-raw/dados-espaciais/uf_sf.shp")) |>
  dplyr::select(cod_stt, geometry)  |>
  dplyr::mutate(cod_stt = as.character(cod_stt))

df_sifilis_uf_sf <- df_sifilis_uf |>
  dplyr::left_join(uf_sf, by = c("uf_res" = "cod_stt")) |>
  sf::st_as_sf()

df_sifilis_uf_sf <- sf::st_transform(df_sifilis_uf_sf, crs = '+proj=longlat +datum=WGS84')
```

3.3.1.2.1 Gráfico

```r
df_sifilis_uf_sf |>
  ggplot2::ggplot() +
  ggplot2::geom_sf(aes(fill = taxa_sifilis)) +
  ggplot2::labs(title = "Taxa de casos de Sífilis por UF",
       fill = "Taxa de sífilis (1000*casos/nascidos)") +
  ggplot2::theme_void() +
  ggplot2::theme(legend.position = "bottom") +
  ggplot2::scale_fill_gradient2(low = "#F2D4F3", high = "#791A78") +
  ggplot2::guides(fill = ggplot2::guide_legend(theme = ggplot2::theme(
    legend.title = ggplot2::element_text(size = 15, face = "bold", colour = "black")
  )))
```



3.3.1.2.2 Tabela

```
df_sifilis_uf |>
  dplyr::select(uf_nome, n_casos_sifg, n_nascidos, taxa_sifilis) |>
  dplyr::mutate(taxa_sifilis = round(taxa_sifilis, 2)) |>
  gt::gt() |>
  gt::cols_label(uf_nome = "UF", n_casos_sifg = "Nº de casos", n_nascidos = "Nº de nascidos"
```

UF	Nº de casos	Nº de nascidos	Taxa de sífilis (1000*n_casos/n_nascidos)
Rondônia	4343	446826	9.72
Acre	5080	277212	18.33
Amazonas	16379	1296500	12.63
Roraima	2530	196850	12.85
Pará	25787	2367885	10.89
Amapá	3703	254570	14.55
Tocantins	5279	415201	12.71
Maranhão	13967	1945302	7.18
Piauí	5715	814743	7.01
Ceará	20114	2144397	9.38
Rio Grande do Norte	7469	781490	9.56
Paraíba	7250	973873	7.44
Pernambuco	25772	2297827	11.22
Alagoas	7759	877218	8.85
Sergipe	7535	565928	13.31

Bahia	34935	3427248	10.19
Minas Gerais	41703	4317864	9.66
Espírito Santo	13545	907282	14.93
Rio de Janeiro	95510	3608836	26.47
São Paulo	132280	9977799	13.26
Paraná	28893	2569794	11.24
Santa Catarina	20194	1569206	12.87
Rio Grande do Sul	41430	2295862	18.05
Mato Grosso do Sul	15640	712616	21.95
Mato Grosso	9258	920071	10.06
Goiás	20277	1576622	12.86
Distrito Federal	6574	718664	9.15

```
df_categoric_av <- df_categoric |>
  dplyr::filter(n_distinct <= 20)

vars_av <- df_categoric_av$variavel
i <- vars_av[1]

## Removendo NAs para contagens de categorias
df_sifilis_categoric_vars <- df_sifilis_categoric_vars |>
  dplyr::mutate(across(where(is.character), ~ ifelse(is.na(.x), "NA", .x))) |>
  dplyr::mutate(ano_diag = ifelse(ano_diag == "****", "NA", ano_diag))

df_vars_categ <- data.frame(variavel = character(), categorias = character())
for(i in vars_av){
  categ <- unique(df_sifilis_categoric_vars[[i]])
  cat("\n")
  table <- df_sifilis_categoric_vars[[i]] |>
    dplyr::as_tibble() |>
    dplyr::group_by(value) |>
    dplyr::summarise(n = n()) |>
    dplyr::ungroup() |>
    dplyr::rename(categoria = value)

  total <- sum(table$n)

  table <- table |>
    dplyr::mutate(perc = round(n/total*100, 0)) |>
```

```
    dplyr::mutate(row = paste0("**", categoria, ":** ", "</b>", n, " (", perc, "%)")) |>
    dplyr::select(row)

  row <- paste0(table$row, collapse = "; \n")
  # table |> gt::gt()
    # gt::cols_label(categoria = "Categoria", n = "Nº de casos")

  df_vars_categ <- rbind(df_vars_categ, data.frame(variavel = i, categorias = row))

  # table_print <- gt::gt(table, rowname_col = "row", groupname_col = "group")
  # print(table_print)
  # cat("\n")
}
```

```
df_vars_categ |>
  gt::gt() |>
  gt::cols_label(variavel = "Variável", categorias = "Categorias: N na categoria (% na categ
  gt::fmt_markdown(columns = everything(), rows = everything(), md_engine = "markdown")
```

Variável	Categorias: N na categoria (% na categoria)
ano_diag	**2007:** 35394 (1%); **2008:** 35798 (1%); **2009:** 36477 (1%); **2010:** 46199 (1%); **2011:** 90393 (2%)
faixa_etaria	**10-14:** 135546 (4%); **15-19:** 665798 (18%); **20-24:** 847379 (22%); **25-29:** 858730 (23%); **30-39:**
raca	**Amarela:** 69670 (2%); **Branca:** 1128091 (30%); **Em branco/Inválido:** 448086 (12%); **Indige**
escolaridade	**1ª a 4ª série incompleta do EF:** 168375 (4%); **4ª série completa do EF:** 307110 (8%); **5ª**

3.3.2 Nascidos com anomalia

```
df_sifilis_anomalia <- df_sifilis |>
  dplyr::group_by(ano_diag) |>
  dplyr::summarise(n_casos_sifg = sum(n_casos_sifg),
                   n_nascidos_anomalia = sum(nasc_c_anom),
                   n_nascidos_vivos = sum(nasc_viv)) |>
  dplyr::ungroup() |>
  dplyr::mutate(taxa_anomalia = n_nascidos_anomalia / n_nascidos_vivos * 1000)
```

3.3.2.1 Tabela

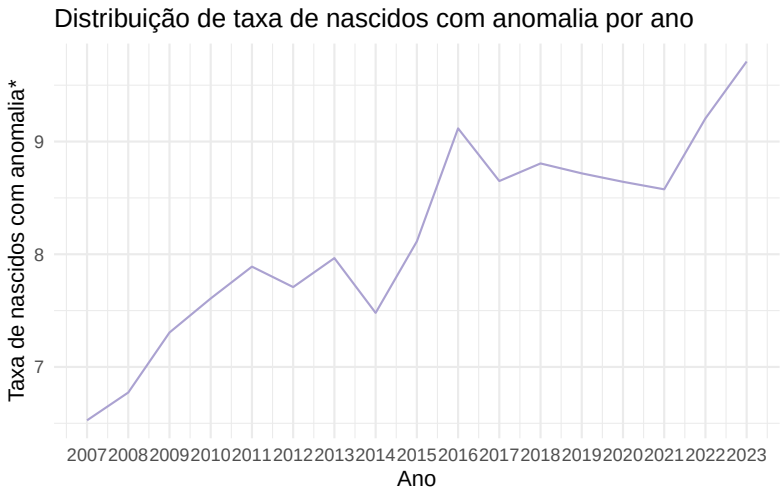
```
df_sifilis_anomalia |>
  dplyr::select(ano_diag, n_nascidos_anomalia, taxa_anomalia, n_casos_sifg, n_nascidos_vivos
  gt::gt() |>
  gt::cols_label(ano_diag = "Ano", n_nascidos_anomalia = "Nº de nascidos com anomalia", taxa
```

Ano	Nº de nascidos com anomalia	Taxa de nascidos com anomalia	Nº de casos de Sífilis	Nº de nascidos
****	0	NaN	48	
2007	18870	6.526413	6259	289
2008	19878	6.773140	7303	293
2009	21051	7.305365	8381	288
2010	21772	7.607619	10075	286
2011	22985	7.890058	13761	291
2012	22400	7.708750	16446	290
2013	23133	7.965835	20922	290
2014	22284	7.479712	26631	297
2015	24485	8.113881	32793	301
2016	26054	9.116803	38317	285
2017	25287	8.649460	49862	292
2018	25932	8.805636	63440	294
2019	24838	8.717700	64637	284
2020	23596	8.642764	66104	273
2021	22959	8.576068	75168	267
2022	23583	9.205198	83033	256
2023	23533	9.709948	35741	242

3.3.2.2 Gráfico

```
library(ggplot2)
df_sifilis_anomalia |>
  dplyr::filter(ano_diag != "****") |>
  dplyr::mutate(ano_diag = as.numeric(ano_diag)) |>
  ggplot2::ggplot() +
  ggplot2::geom_line(aes(x = ano_diag, y = taxa_anomalia), color = "#ABA2D1") +
  ggplot2::labs(title = "Distribuição de taxa de nascidos com anomalia por ano",
       x = "Ano",
       y = "Taxa de nascidos com anomalia*") +
  ggplot2::theme_minimal() +
  ggplot2::scale_y_continuous(labels = scales::number) +
```

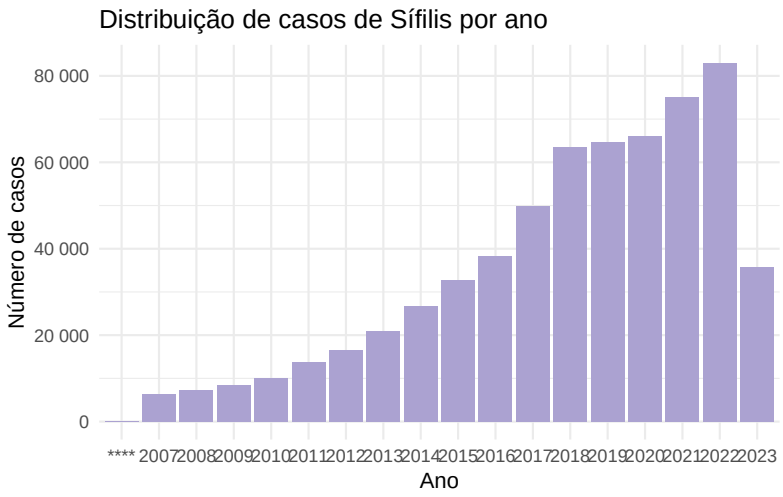
```
  ggplot2::scale_x_continuous(breaks = seq(2007, 2023, 1))
```



4 Análise bivariada

4.1 Variáveis X por ano

```
df_sifilis_biv_ano <- df_sifilis |>
  dplyr::group_by(ano_diag) |>
  dplyr::summarise(n_casos_sifg = sum(n_casos_sifg)) |>
  dplyr::ungroup()

df_sifilis_biv_ano |>
  ggplot2::ggplot() +
  ggplot2::geom_col(aes(x = ano_diag, y = n_casos_sifg), fill = "#ABA2D1") +
  ggplot2::labs(title = "Distribuição de casos de Sífilis por ano",
       x = "Ano",
       y = "Número de casos") +
  ggplot2::theme_minimal() +
  ggplot2::scale_y_continuous(labels = scales::number)
```


Distribuição de casos de Sífilis por ano
Número de casos
80 000
60 000
40 000
20 000
0
**** 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023
Ano