

# Atividade Avaliativa

## Análise Estatística para Data Science

Ana Maria Alves da Silva

2024-09-14

### Questão 1

Utilizando os comandos apresentados em aula e o banco de dados *dados\_saude* determine os eventos:

- A: a população indígena da região tem mais que 5000 e menos que 10000 habitantes;
- B: o percentual (ou proporção) da população indígena que tomou a segunda dose ou dose única é maior que 75% (ou 0.75) da população;

Calcule:  $P(A)$ ,  $P(B)$ ,  $P(A \cup B)$  e  $P(A \cap B)$ .

### Solução:

1. Carregamento dos dados

Antes de responder aos itens solicitados, vamos realizar uma análise inicial dos dados.

```
df_saude <- read_excel(
  "/Users/anamaria/especializacao/modulo_5/atividade/dados_saude.xlsx")
print(df_saude)
```

```
## # A tibble: 34 x 13
##   DSEI      PI    D1    P1 `D2&DU`    P2 D1_3a4A D1_5a17A D1_18M D2_3a4A
##   <chr>    <dbl> <dbl> <dbl>    <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 Alagoas e se~ 12818 12627 0.985    12283 0.958      481      3615      8531      426
## 2 Altamira      4383  4286 0.978      4047 0.923      271      1822      2193      176
## 3 Alto rio jur~ 16361 12899 0.788      9959 0.609      692      5086      7121      482
## 4 Alto rio neg~ 24789 24789 1          23771 0.959     1706      7003     16080      688
## 5 Alto rio pur~ 10968 10157 0.926      7773 0.709      639      4021      5497      193
## 6 Alto rio sol~ 65706 54793 0.834     37678 0.573     1596     23332     29865      452
## 7 Amapá e nort~ 11311  9770 0.864      7116 0.629      383      3743      5644      111
## 8 Araguaia      4992  3251 0.651      2213 0.443        0      1186      2065        0
## 9 Bahia        31927 30741 0.963     29258 0.916     1039      8908     20794      785
## 10 Ceará       27697 27620 0.997     27532 0.994      986      6322     20312      986
## # i 24 more rows
## # i 3 more variables: D2_5a17A <dbl>, D2_18M <dbl>, REF1_18M <dbl>
```

Supondo que as colunas do dataframe seguem o mesmo padrão das colunas do mesmo dataframe vistas em aulas, temos:

- Colunas no conjunto de dados:

C1 : DSEI (Distritos Sanitários Especiais Indígenas);  
C2: População Indígena (nº de indígenas por distrito);  
C3: Dose 1 (nº de indígenas vacinados com a 1a dose);  
C4: % Dose 1 (percentual de indígenas vacinados com a 1a dose);  
C5: Dose 2 e Única (nº de indígenas vacinados com a 2a dose e dose única);  
C6: % Dose 2 e Única (percentual de indígenas vacinados com a 2a dose e dose única);  
C7: Dose 1 (3 a 4 anos): (nº de indígenas vacinados com a 1a dose com faixa etária entre 3 e 4 anos);  
C8: Dose 1 (5 a 17 anos): (nº de indígenas vacinados com a 1a dose com faixa etária entre 5 e 17 anos);  
C9: Dose 1 (>18 anos): (nº de indígenas vacinados com a 1a dose com faixa etária maior que 18 anos);  
C10: Dose 2 (3 a 4 anos): (nº de indígenas vacinados com a 2a dose com faixa etária entre 3 e 4 anos);  
C11: Dose 2 (5 a 17 anos): (nº de indígenas vacinados com a 2a dose com faixa etária entre 5 e 17 anos);  
C12: Dose 2 (>18 anos): (nº de indígenas vacinados com a 2a dose com faixa etária maior que 18 anos);  
C13: Reforço 1 (> 18 anos): (nº de indígenas vacinados com a 1a dose de Reforço com faixa etária maior que 18 anos);

- Renomeando as colunas para melhor entedimento

```
colnames(df_saude) <- c('DSEI', 'Populacao_Indigena', 'Dose1', 'Percentual_Dose1',  
                        'Dose2_Unica', 'Percentual_Dose2_Unica', 'Dose1_3a4',  
                        'Dose1_5a17', 'Dose1_18+', 'Dose2_3a4', 'Dose2_5a17',  
                        'Dose2_18+', 'Reforco_18+')
```

- Tamanho do conjunto de dados:

```
print(dim(df_saude))
```

```
## [1] 34 13
```

- Verificando se é um dataframe:

```
is_data_frame <- is.data.frame(df_saude)  
print(is_data_frame)
```

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

- Verificando se há dados duplicados:

```
duplicados <- duplicated(df_saude)
ha_duplicados <- any(duplicados)
print(ha_duplicados)
```

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

Logo, nosso dataframe tem 13 colunas e 34 linhas. Além disso, não há linhas duplicadas.

- Visualizando um resumo dos dados:

```
summary_df <- summary(df_saude)
print(summary_df)
```

```
##      DSEI      Populacao_Indigena      Dose1      Percentual_Dose1
## Length:34      Min.   : 4383      Min.   : 3251      Min.   :0.5559
## Class :character 1st Qu.: 9172      1st Qu.: 7847      1st Qu.:0.8335
## Mode  :character Median :15826      Median :13664      Median :0.9295
##              Mean   :20816      Mean   :18588      Mean   :0.8837
##              3rd Qu.:27677      3rd Qu.:26912      3rd Qu.:0.9719
##              Max.   :75542      Max.   :67751      Max.   :1.0000
## Dose2_Unica Percentual_Dose2_Unica Dose1_3a4      Dose1_5a17
## Min.   : 1985      Min.   :0.3390      Min.   : 0.0      Min.   : 1046
## 1st Qu.: 6234      1st Qu.:0.6315      1st Qu.: 273.5      1st Qu.: 2630
## Median :12418      Median :0.8206      Median : 625.0      Median : 5144
## Mean   :16224      Mean   :0.7684      Mean   : 749.4      Mean   : 6534
## 3rd Qu.:25857      3rd Qu.:0.9160      3rd Qu.:1059.2      3rd Qu.: 8230
## Max.   :58074      Max.   :0.9940      Max.   :2184.0      Max.   :24348
## Dose1_18+      Dose2_3a4      Dose2_5a17      Dose2_18+
## Min.   : 2065      Min.   : 0.0      Min.   : 192      Min.   : 1633
## 1st Qu.: 4931      1st Qu.: 135.0      1st Qu.: 1936      1st Qu.: 4453
## Median : 8526      Median : 311.0      Median : 3786      Median : 8409
## Mean   :11305      Mean   : 460.5      Mean   : 4985      Mean   :10779
## 3rd Qu.:16130      3rd Qu.: 759.0      3rd Qu.: 7802      3rd Qu.:16029
## Max.   :41441      Max.   :1536.0      Max.   :16981      Max.   :40022
## Reforco_18+
## Min.   : 749
## 1st Qu.: 3209
## Median : 6304
## Mean   : 9013
## 3rd Qu.:14170
## Max.   :31495
```

## 2. Determinando os eventos A e B

```
# Evento A: população indígena entre 5000 e 10000
evento_A <- df_saude$Populacao_Indigena > 5000 & df_saude$Populacao_Indigena < 10000

# Evento B: percentual de segunda dose maior que 75%
evento_B <- df_saude$Percentual_Dose2_Unica > 0.75
```

## 3. Determinando as Probabilidades

```

# População indígena total
populacao_total <- sum(df_saude$Populacao_Indigena)

# P(A)
P_A <- round(sum(df_saude$Populacao_Indigena[evento_A]) / populacao_total, 4)*100

# P(B)
P_B <- round(sum(df_saude$Populacao_Indigena[evento_B]) / populacao_total, 4)*100

# P(A ∪ B)
P_A_union_B <- round(
  sum(
    df_saude$Populacao_Indigena[evento_A | evento_B]
  ) / populacao_total, 4
)*100

# P(A ∩ B)
P_A_intersection_B <- round(
  sum(
    df_saude$Populacao_Indigena[evento_A & evento_B]
  ) / populacao_total, 4
)*100

cat("Probabilidade do evento A: ", P_A, "%", sep = "", "\n")

```

```
## Probabilidade do evento A: 5.89%
```

```
cat("Probabilidade do evento B: ", P_B, "%", sep = "", "\n")
```

```
## Probabilidade do evento B: 62.3%
```

```
cat("Probabilidade da união de A e B: ", P_A_union_B, "%", sep = "", "\n")
```

```
## Probabilidade da união de A e B: 64.29%
```

```
cat("Probabilidade da interseção de A e B: ", P_A_intersection_B, "%", sep = "", "\n")
```

```
## Probabilidade da interseção de A e B: 3.9%
```

## Questão 2

O banco de dados dados\_vacinacao\_sp.xls contém informações sobre a quantidade de vacinas contra COVID-19 aplicadas, diariamente, em 2022, na cidade de São Paulo. Responda:

- obtenha um intervalo de confiança de 90% para a quantidade média de vacinas contra COVID-19 aplicadas, diariamente, em 2022, na cidade de São Paulo. Interprete o resultado.
- pode-se afirmar que a quantidade média diária de vacinas contra COVID-19 aplicadas em São Paulo, no ano de 2022, é igual a 30 mil doses, ao nível de significância de 5%? Ou seja

–  $H_0$  : A média de vacinas aplicadas diariamente em São Paulo em 2022 é igual a 30 mil –  $H_1$ : A média de vacinas aplicadas diariamente em São Paulo em 2022 é diferente de 30 mil Interprete o resultado.

## Solução:

1. Carregamento dos dados

Antes de responder aos itens solicitados, vamos realizar uma análise inicial dos dados.

```
df_vacina_sp <- read_excel(
  "/Users/anamaria/especializacao/modulo_5/atividade/dados_vacinacao_sp.xls")
print(df_vacina_sp)
```

```
## # A tibble: 325 x 2
##   `Data da Vacina`   `Total de Doses Aplicadas`
##   <dtm>                <dbl>
## 1 2022-01-01 00:00:00           34
## 2 2022-01-02 00:00:00          2780
## 3 2022-01-03 00:00:00         88565
## 4 2022-01-04 00:00:00         86571
## 5 2022-01-05 00:00:00        93133
## 6 2022-01-06 00:00:00        90036
## 7 2022-01-07 00:00:00        87219
## 8 2022-01-08 00:00:00        60256
## 9 2022-01-09 00:00:00         2506
## 10 2022-02-19 00:00:00        51136
## # i 315 more rows
```

- Tamanho do conjunto de dados:

```
print(dim(df_vacina_sp))
```

```
## [1] 325  2
```

- Verificando se é um dataframe:

```
is_data_frame_2 <- is.data.frame(df_vacina_sp)
print(is_data_frame_2)
```

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

- Verificando se há dados duplicados:

```
duplicados <- duplicated(df_vacina_sp)
ha_duplicados <- any(duplicados)
print(ha_duplicados)
```

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

- Visualizando um resumo dos dados:

```
summary_df_sp <- summary(df_vacina_sp)
print(summary_df_sp)
```

```
## Data da Vacina                Total de Doses Aplicadas
## Min.      :2022-01-01 00:00:00.00 Min.      :    1
## 1st Qu.:2022-05-02 00:00:00.00 1st Qu.:  8610
## Median :2022-07-22 00:00:00.00 Median : 20483
## Mean    :2022-07-20 21:24:55.38 Mean    : 28607
## 3rd Qu.:2022-10-11 00:00:00.00 3rd Qu.: 47212
## Max.    :2022-12-31 00:00:00.00 Max.     :105884
```

## 2. Cálculo da Média, Desvio Padrão

```
media_doses <- mean(df_vacina_sp$`Total de Doses Aplicadas`)
desvio_padrao <- sd(df_vacina_sp$`Total de Doses Aplicadas`)
length_amostra <- nrow(df_vacina_sp)
cat("Média", media_doses, "\n")
```

```
## Média 28607.11
```

```
cat("Desvio Padrão", desvio_padrao, "\n")
```

```
## Desvio Padrão 25867.07
```

```
cat("Tamanho da Amostra", length_amostra, "\n")
```

```
## Tamanho da Amostra 325
```

3. Item(a) obtenha um intervalo de confiança de 90% para a quantidade média de vacinas contra COVID-19 aplicadas, diariamente, em 2022, na cidade de São Paulo. Para isso usaremos a função *t.test*.

```
intervalo_conf <- t.test(df_vacina_sp$`Total de Doses Aplicadas`, conf.level = 0.90)
print(intervalo_conf$conf.int)
```

```
## [1] 26240.23 30973.99
## attr(,"conf.level")
## [1] 0.9
```

Com o resultado acima, estamos 90% confiantes de que a média diária verdadeira de vacinas aplicadas em São Paulo durante 2022 está entre 26.240,23 e 30.973,99 doses.

4. Item (b) pode-se afirmar que a quantidade média diária de vacinas contra COVID-19 aplicadas em São Paulo, no ano de 2022, é igual a 30 mil doses, ao nível de significância de 5%?

```
# Definir a média hipotética
media_hipotetica <- 30000

# Realizar o teste t
teste_t <- t.test(df_vacina_sp$`Total de Doses Aplicadas`, mu = media_hipotetica)

# Exibir o resultado do teste t
print(teste_t)
```

```
##
## One Sample t-test
##
## data: df_vacina_sp$`Total de Doses Aplicadas`
## t = -0.97076, df = 324, p-value = 0.3324
## alternative hypothesis: true mean is not equal to 30000
## 95 percent confidence interval:
## 25784.32 31429.90
## sample estimates:
## mean of x
## 28607.11

# Interpretar o p-valor
if(teste_t$p.value < 0.05) {
  cat(
    "Rejeitamos a hipótese nula, H0. A média diária de vacinas é diferente de 30.000 doses.\n")
} else {
  cat(
    "Não rejeitamos a hipótese nula, H0. Não há evidências suficientes para afirmar\n
    que a média diária de vacinas é diferente de 30.000 doses.\n")
}

## Não rejeitamos a hipótese nula, H0. Não há evidências suficientes para afirmar
##
## que a média diária de vacinas é diferente de 30.000 doses.
```

A média diária observada é de 28.607 doses. O valor do teste t é -0,97, e o valor p é 0,332.

Como o valor p é maior que o nível de significância de 5% (0,05), não rejeitamos a hipótese nula H0. Isso significa que não temos evidências suficientes para afirmar que a quantidade média diária de vacinas aplicadas foi diferente de 30.000 doses.

### Questão 3

O Departamento de Monitoramento e Avaliação (DEMAS) da Secretaria de Informação e Saúde Digital (SEIDIGI) desenvolveu em parceria com o Departamento de Imunização e Doenças Imunopreveníveis (DIMU) da Secretaria de Vigilância em Saúde e Ambiente (SVSA), um painel com dados sobre os imunizantes de COVID-19, clique aqui.

As seguintes informações foram obtidas em 06/09/2024:

- Quilombolas vacinados com a primeira dose no Brasil: 604.329
- População quilombola no Brasil: 1.133.106
- Quilombolas vacinados com a primeira dose na região norte: 79.111
- População quilombola na região norte: 154.911

Responda:

- (a) é possível afirmar que a proporção de quilombolas vacinados com a primeira dose na região norte do Brasil é igual à proporção nacional? Considere o nível de significância de 5%. Interprete o resultado.
- (b) obtenha um intervalo de confiança de 99% para a proporção de quilombolas vacinados no Brasil. Interprete o resultado.

Obs.: para o cálculo das proporções de quilombolas vacinados na região norte ou no país, utilize 3 casas decimais.

## Solução:

### 1. Carregamento dos dados

Resalto que para essa questão os dados não foram fornecidos, extraí os dados do link informado no dia 11/09 às 21h. Logo, pode haver divergência nas informações. Nesse período houve atualização dos dados.

Antes de responder aos itens solicitados, vamos realizar uma análise inicial dos dados.

```
df_pq <- read_excel(
  "/Users/anamaria/especializacao/modulo_5/atividade/dados_populacao_quilombola.xlsx")
print(df_pq)
```

```
## # A tibble: 166,365 x 13
##   Município      cod_ibge UF   Região Fabricante      Data
##   <chr>          <dbl> <chr> <chr> <chr>      <dtm>
## 1 Acrelândia      120001 AC   Norte University of Oxfo~ 2021-07-12 00:00:00
## 2 Bujari          120013 AC   Norte BioNTech/Fosun Pha~ 2022-08-11 00:00:00
## 3 Bujari          120013 AC   Norte University of Oxfo~ 2021-09-20 00:00:00
## 4 Cruzeiro Do Sul 120020 AC   Norte BioNTech/Fosun Pha~ 2022-11-22 00:00:00
## 5 Cruzeiro Do Sul 120020 AC   Norte University of Oxfo~ 2022-07-15 00:00:00
## 6 Rio Branco      120040 AC   Norte BioNTech/Fosun Pha~ 2022-02-09 00:00:00
## 7 Rio Branco      120040 AC   Norte BioNTech/Fosun Pha~ 2022-04-02 00:00:00
## 8 Rio Branco      120040 AC   Norte JANSSEN-CILAG      2021-12-20 00:00:00
## 9 Rio Branco      120040 AC   Norte University of Oxfo~ 2021-11-03 00:00:00
## 10 Rio Branco     120040 AC   Norte University of Oxfo~ 2023-01-18 00:00:00
## # i 166,355 more rows
## # i 7 more variables: `Total de doses aplicadas` <dbl>,
## #   `Quilombolas Vacinados (Dose 1)` <dbl>,
## #   `Quilombolas Vacinados (Dose 2)` <dbl>,
## #   `Quilombolas Vacinados (Única)` <dbl>,
## #   `Quilombolas Vacinados (Adicional)` <dbl>,
## #   `Quilombolas Vacinados (Reforço)` <dbl>, ...
```