

# Tarefa IV de survey

Felipe Lamarca

2025-04-07

Nesta tarefa, aplico a técnica de *Multilevel Regression with Poststratification* (MrP) a um survey online com 2.015 adultos brasileiros recrutados via Facebook, com o objetivo de recuperar estimativas populacionais a partir de uma amostra enviesada. O procedimento envolve, inicialmente, o ajuste de modelos de regressão multinível para prever atitudes e comportamentos em estratos sociodemográficos relevantes, seguido da pós-estratificação com base na PNAD Contínua. Testo duas especificações de MrP para estimar o voto em Jair Bolsonaro no 1º turno de 2018, usando o valor oficial como referência. O modelo com interceptos e *slopes* variáveis por estado mostrou melhor desempenho, convergindo para a proporção real. Com essa configuração, estimo que 67% da população se opõe à legalização do aborto, resultado consistente com pesquisas relativamente recentes de dois importantes institutos de pesquisa de opinião brasileiros. Os achados demonstram o potencial do MrP para corrigir vieses na amostra e gerar estimativas confiáveis, mesmo a partir de surveys não-probabilísticos. A qualidade dos resultados, no entanto, é altamente dependente de uma boa especificação do modelo, o que depende de um bom conhecimento de negócio e uma análise exploratória detalhada.

## Introdução

A técnica de *Multilevel Regression with Poststratification (MRP)* combina dois passos complementares. No primeiro, ajusta-se um modelo de regressão multinível que estima, para cada estrato sociodemográfico relevante (UF, sexo, idade, cor/raça, escolaridade etc.), a probabilidade de o respondente exibir o comportamento ou atitude de interesse. No segundo passo, essas probabilidades são pós-estratificadas: multiplicamos cada predição pelo tamanho real de seu estrato na população (aqui, a PNAD Contínua), gerando uma estimativa ponderada que corrige as distorções da amostra original.

A regressão multinível é uma técnica estatística que permite modelar dados com estrutura hierárquica ou agrupada – como indivíduos dentro de estados, partidos, ou outras unidades geográficas ou sociais Gelman e Hill (2007). Esse modelo incorpora variações em diferentes níveis ao permitir que interceptos (e, se necessário, coeficientes) variem entre grupos, ao invés de assumir que todos os grupos compartilham os mesmos parâmetros. Essa abordagem é especialmente poderosa porque melhora a precisão das estimativas em subgrupos com poucos dados, utilizando o conceito de “shrinkage”, que suaviza estimativas extremas com base na média geral, equilibrando variabilidade e robustez.

Aplicamos esse procedimento a um survey *online* com 2.015 adultos brasileiros recrutados via Facebook Smith e Boas (2024). Primeiro, diagnosticamos os vieses da amostra comparando-a à estrutura populacional. Em seguida, faço uma análise exploratória – conforme sugerido por Ghitza e Gelman

(2013) – não-exaustiva para, finalmente, testar duas especificações de MrP para estimar a proporção de voto em Jair Bolsonaro no 1º turno de 2018 – cujo valor, conhecido, é de 46.03%, útil como *ground truth*. O modelo mais simples, apenas com interceptos variáveis, subestimou o apoio. O segundo, incluindo *slopes* variáveis para idade e escolaridade por estado, convergiu para a proporção verdadeira.

Com essa especificação mais robusta projetamos, então, a proporção da população contrária à legalização do aborto. O MrP indica que cerca de 67% dos brasileiros se declaram contra, resultado alinhado às pesquisas publicadas pela Datafolha (2024) e Quaest (2023) há não muito tempo atrás. Regionalmente, a oposição é maior no Sul e Centro-Oeste e menor no Norte. Em geral, o exercício mostra como, mesmo partindo de um survey não-probabilístico e enviesado, o MrP permite recuperar estimativas nacionais próximas da realidade demográfica.

## Setup da tarefa e manipulação dos dados

Aqui, simplesmente faço a importação das bibliotecas necessárias à análise, defino uma seed e faço a leitura dos dados. Note que extraio também dados do pacote `geobr` para possibilitar a apresentação de algumas análises exploratórias na forma de mapas. No mais, crio uma coluna nos dois bancos de dados principais derivando a região à qual cada estado pertence.

```
# bibliotecas necessarias
library(tidyverse)
library(geobr)
library(sf)
library(scales)
library(rlang)
library(lme4)
library(gt)

# seed (modelos multinivel sao ajustados por algoritmos de otimizacao)
set.seed(42)

# leitura dos bancos -- estratos e o survey
load("data/estratos.Rda")
df <- read.csv("data smith_boas_2019.csv")

# Baixa os dados geográficos das UFs
ufs_sf <- geobr::read_state(code_state = "all", year = 2019)

# Mapeamento de UF para Região
uf_para_regiao <- c(
  # Norte
  "AC" = "Norte", "AP" = "Norte", "AM" = "Norte", "PA" = "Norte",
  "RO" = "Norte", "RR" = "Norte", "TO" = "Norte",

  # Nordeste
  "MA" = "Nordeste", "PI" = "Nordeste", "CE" = "Nordeste", "RN" = "Nordeste",
  "PB" = "Nordeste", "PE" = "Nordeste", "AL" = "Nordeste", "SE" = "Nordeste",
```

```

"BA" = "Nordeste",
# Sudeste
"MG" = "Sudeste", "ES" = "Sudeste", "RJ" = "Sudeste", "SP" = "Sudeste",
# Sul
"PR" = "Sul", "SC" = "Sul", "RS" = "Sul",
# Centro-Oeste
"MT" = "Centro-Oeste", "MS" = "Centro-Oeste", "GO" = "Centro-Oeste", "DF" = "Centro-Oeste"
)

# Adiciona a coluna 'regiao' aos dataframes
estratos$regiao <- uf_para_regiao[estratos$uf]
df$regiao <- uf_para_regiao[df$uf]

```

## Metodologia

### Comparação entre a amostra e a população

De saída, vale realizar uma análise preliminar do perfil da amostra coletada em relação à população brasileira. O gráfico abaixo mostra a diferença entre os percentuais da amostra e da população em cada estado – quanto mais azul for o estado, mais sobre-representada está o perfil da amostra naquele estado em relação à população; ao contrário, quanto mais vermelho, mais sub-representado está esse estado na amostra.

```

# Proporção por UF no survey
pop_por_uf_survey <- df %>%
  group_by(uf) %>%
  summarise(n_uf_survey = n()) %>%
  mutate(prop_survey = n_uf_survey / sum(n_uf_survey))

# Proporção por UF na PNAD
pop_por_uf_pnad <- estratos %>%
  group_by(uf) %>%
  summarise(n_uf_pnad = sum(n)) %>%
  mutate(prop_pnad = n_uf_pnad / sum(n_uf_pnad))

# Junta as duas fontes
comparativo <- left_join(pop_por_uf_survey, pop_por_uf_pnad, by = "uf") %>%
  mutate(
    dif = (prop_survey - prop_pnad) * 100
  )

# Faz o join com a geometria
mapa <- ufs_sf %>%

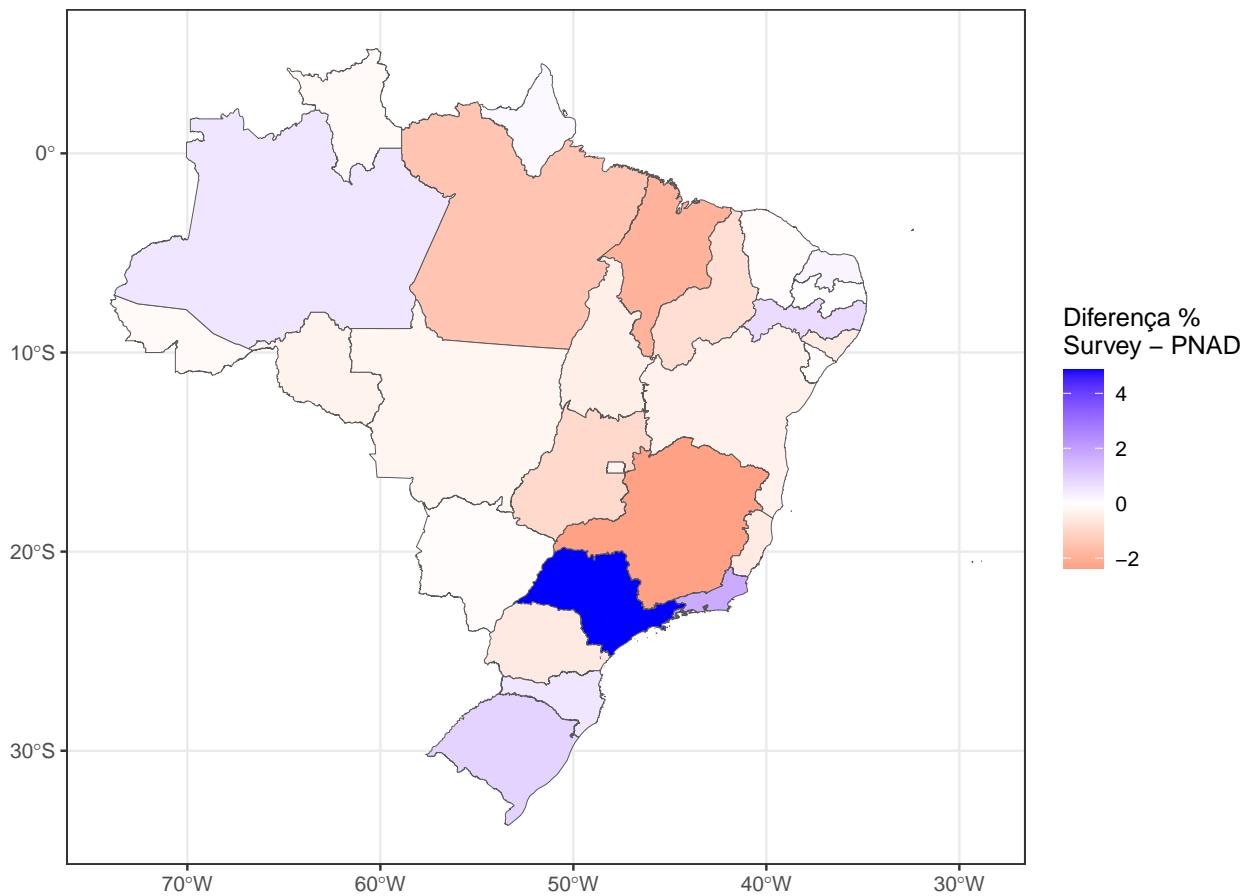
```

```
left_join(comparativo, by = c("abbrev_state" = "uf"))

# Plota o mapa com a razão
ggplot(mapa) +
  geom_sf(aes(fill = dif)) +
  scale_fill_gradient2(
    low = "red", mid = "white", high = "blue",
    midpoint = 0,
    name = "Diferença %\nSurvey - PNAD"
  ) +
  labs(
    title = "Diferença entre proporção no Survey e na PNAD por UF",
    subtitle = "Valores > 0 indicam sobre-representação no survey",
    fill = NULL
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    plot.subtitle = element_text(size = 12)
  )
)
```

## Diferença entre proporção no Survey e na PNAD por UF

Valores > 0 indicam sobre-representação no survey



Para a maior parte dos estados, as diferenças são sutis. Elas são mais significativas no caso do estado de São Paulo e Minas Gerais. No primeiro caso, a população está sobre-representada em relação à população em aproximadamente 4 pontos percentuais; no segundo, ela está sub-representada em aproximadamente 2 pontos percentuais.

A função abaixo nos possibilita fazer o mesmo tipo de análise para as demais variáveis. Nesse caso, apresento tanto um gráfico de barras simples comparando as distribuições marginais de uma variável de interesse qualquer na amostra e na população, quanto um mapa para cada categoria dessa variável.

```
comparar_variavel <- function(df, estratos, variavel) {  
  var_sym <- sym(variavel)  
  
  # --- Proporção total Brasil (gráfico de barras) ---  
  prop_survey <- df %>%  
    group_by(!!var_sym) %>%  
    summarise(n = n(), .groups = "drop") %>%  
    mutate(prop = n / sum(n), fonte = "survey")  
  
  prop_pnad <- estratos %>%  
    group_by(!!var_sym) %>%
```

```

summarise(n = sum(n), .groups = "drop") %>%
mutate(prop = n / sum(n), fonte = "pnad")

prop_total <- bind_rows(prop_survey, prop_pnad)

grafico_barras <- ggplot(prop_total, aes(x = fonte, y = prop, fill = fonte)) +
  geom_col(position = "dodge", width = 0.7) +
  facet_wrap(as.formula(paste("~", variavel)), ncol = 6) +
  scale_y_continuous(labels = percent_format(accuracy = 1)) +
  labs(
    title = paste("Distribuição por", variavel, ": Survey vs. PNAD"),
    x = NULL,
    y = "Percentual"
  ) +
  theme_bw() +
  theme(
    legend.position = "none",
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12),
    strip.text = element_text(size = 13),
    plot.title = element_text(size = 20, face = "bold")
  )
)

print(grafico_barras)

# --- Mapa facetado por categoria ---
df_map <- df %>%
  group_by(uf, !!var_sym) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by (!!var_sym) %>%
  mutate(prop_survey = n / sum(n)) %>%
  select(-n)

estratos_map <- estratos %>%
  group_by(uf, !!var_sym) %>%
  summarise(n = sum(n), .groups = "drop") %>%
  group_by (!!var_sym) %>%
  mutate(prop_pnad = n / sum(n)) %>%
  select(-n)

comparativo <- full_join(df_map, estratos_map, by = c("uf", as_string(var_sym))) %>%
  mutate(
    dif = (prop_survey - prop_pnad) * 100,
    categoria = as.character (!!var_sym)
  )

mapa_all <- ufs_sf %>%

```

```

left_join(comparativo, by = c("abbrev_state" = "uf"))

# Calcula os centroides para adicionar siglas
centroides <- st_centroid(ufs_sf) %>%
  st_coordinates() %>%
  as_tibble() %>%
  bind_cols(ufs_sf["abbrev_state"])

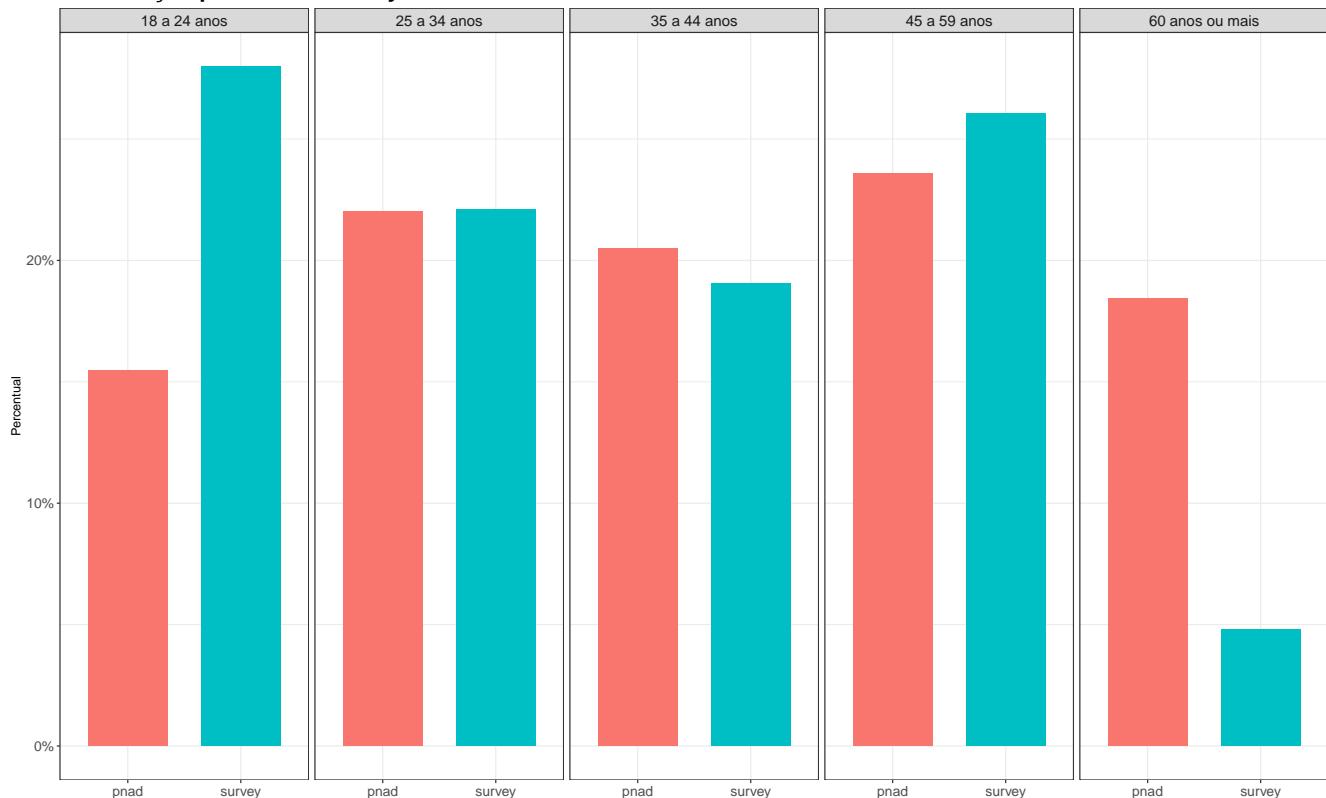
# Mapa com contorno + siglas
p_mapa <- ggplot(mapa_all) +
  geom_sf(aes(fill = dif), color = "gray30", size = 0.2) + # contorno
  geom_text(
    data = centroides,
    aes(X, Y, label = abbrev_state),
    size = 3, color = "black"
  ) +
  scale_fill_gradient2(
    low = "red", mid = "white", high = "blue",
    midpoint = 0,
    name = "Diferença %"
  ) +
  facet_wrap(~ categoria) +
  labs(
    title = paste("Diferença percentual entre Survey e PNAD por UF e categoria de", variavel),
    subtitle = "Valores > 0 indicam sobre-representação no survey"
  ) +
  theme_bw() +
  theme(
    strip.text = element_text(size = 12),
    plot.title = element_text(size = 18, face = "bold"),
    plot.subtitle = element_text(size = 13)
  )
  print(p_mapa)
}

```

Observe os gráficos abaixo, que mostram (i) a distribuição marginal da idade na amostra e na população e (ii) a diferença entre a distribuição na amostra e na população por estado.

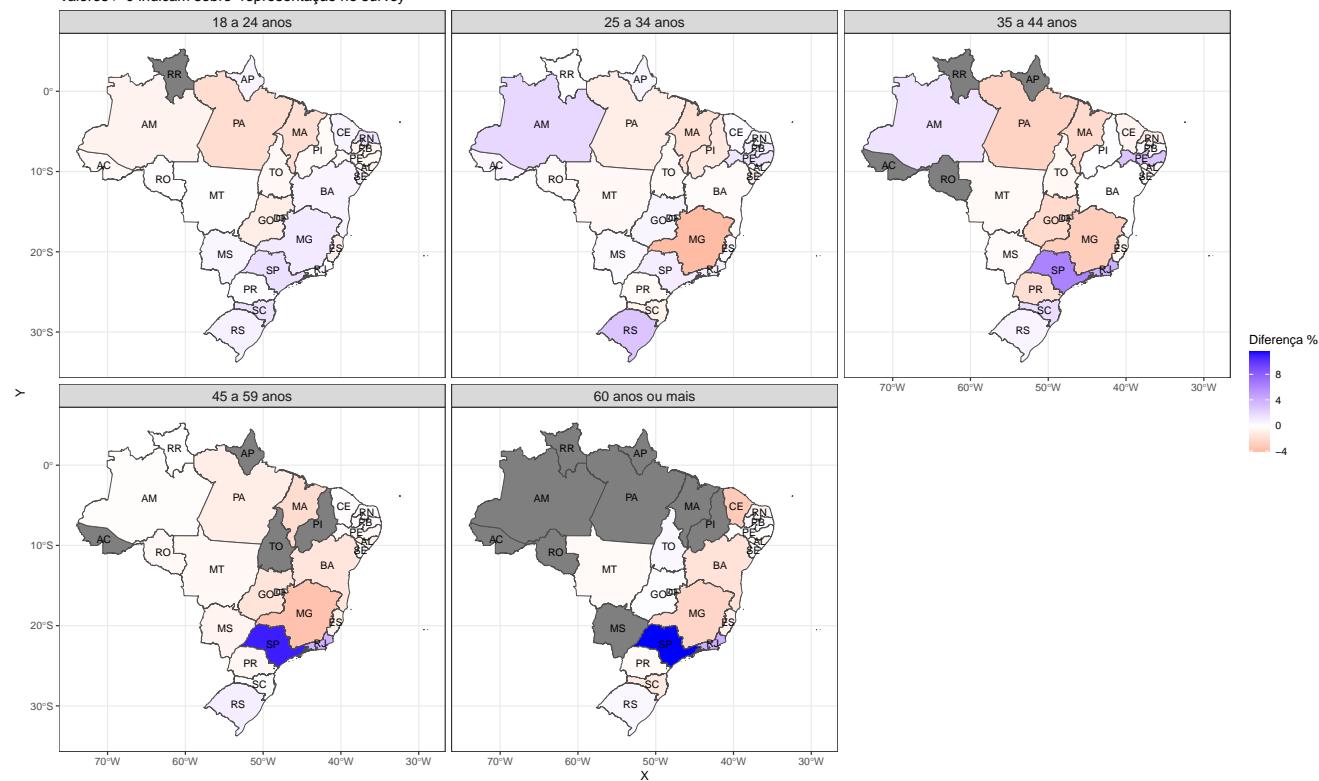
```
comparar_variavel(df, estratos, "idade")
```

### Distribuição por idade : Survey vs. PNAD



### Diferença percentual entre Survey e PNAD por UF e categoria de idade

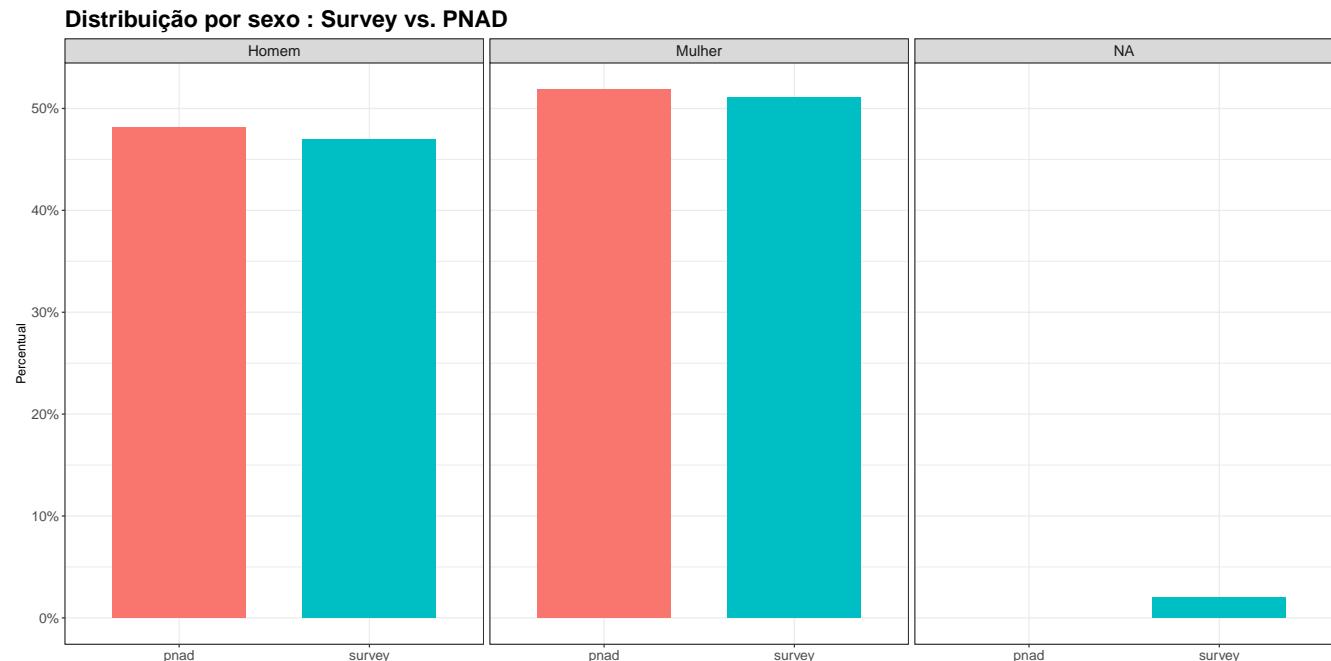
Valores > 0 indicam sobre-representação no survey



Há diferenças substantivas. Em particular, observamos que indivíduos de 18 a 24 anos estão altamente sobre-representados na amostra, enquanto indivíduos de 60 anos ou mais estão fortemente sub-representados. Observando a distribuição por estados, inclusive, é possível verificar que boa parte dos estados do Norte (todos, aliás, com exceção do Tocantins) não possuem representantes idosos na amostra.

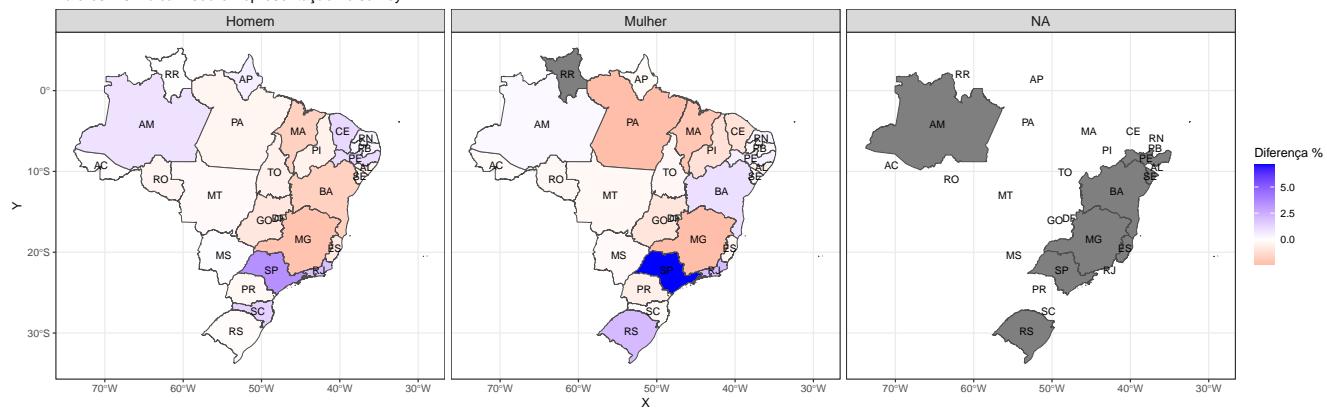
No que diz respeito ao sexo, as diferenças são menores. Note, no entanto, que há alguns poucos valores nulos na base. Quando utilizamos MrP, o impacto negativo desses valores nulos tende a ser minimizado ao incluirmos outras covariáveis no modelo de regressão.

```
comparar_variavel(df, estratos, "sexo")
```



### Diferença percentual entre Survey e PNAD por UF e categoria de sexo

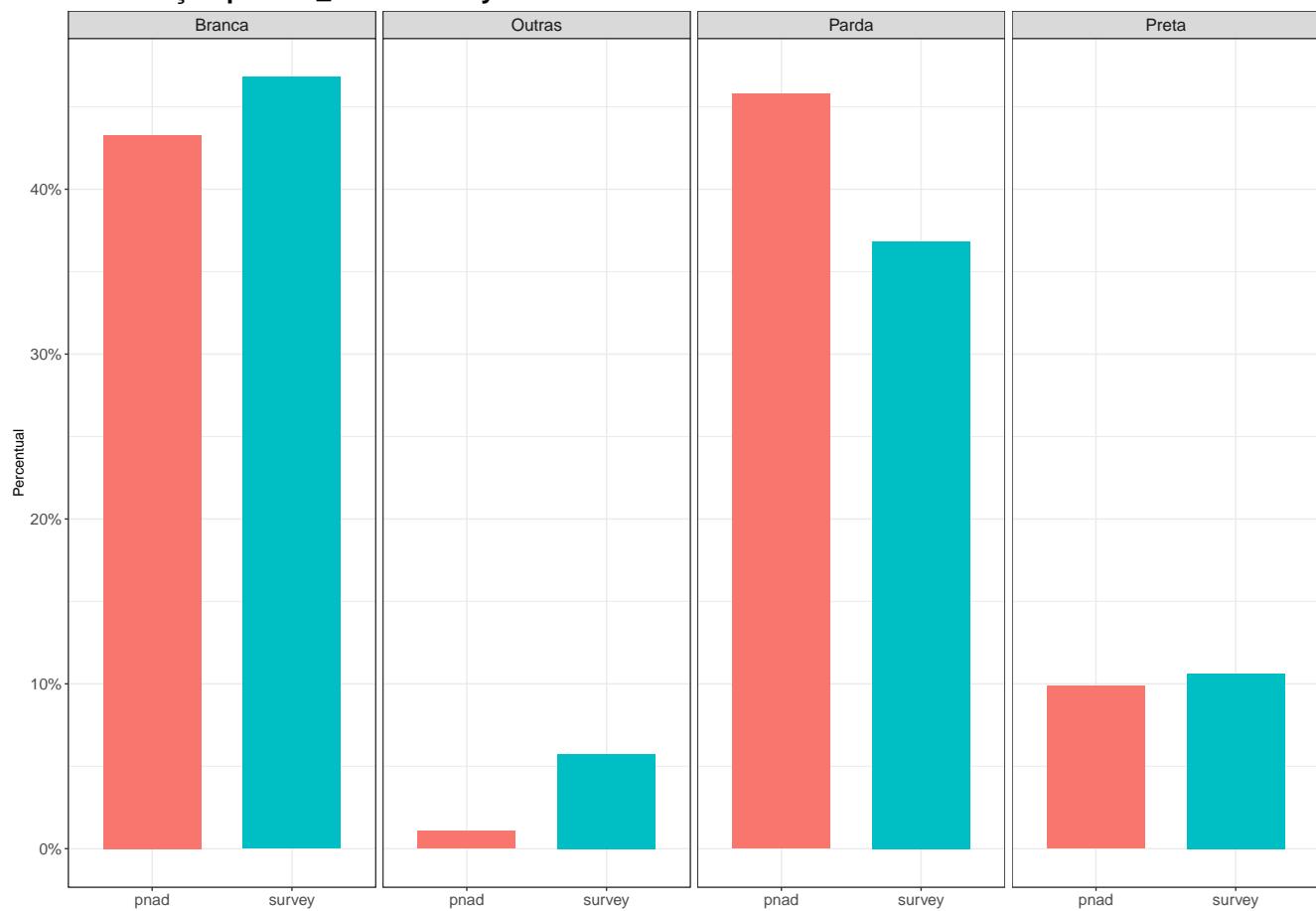
Valores > 0 indicam sobre-representação no survey



Em relação à cor/raça, há diferenças mais significativas especialmente no caso da população parda e “outras”. A primeira é sub-representada no survey e a segunda é razoavelmente sobre-representada.

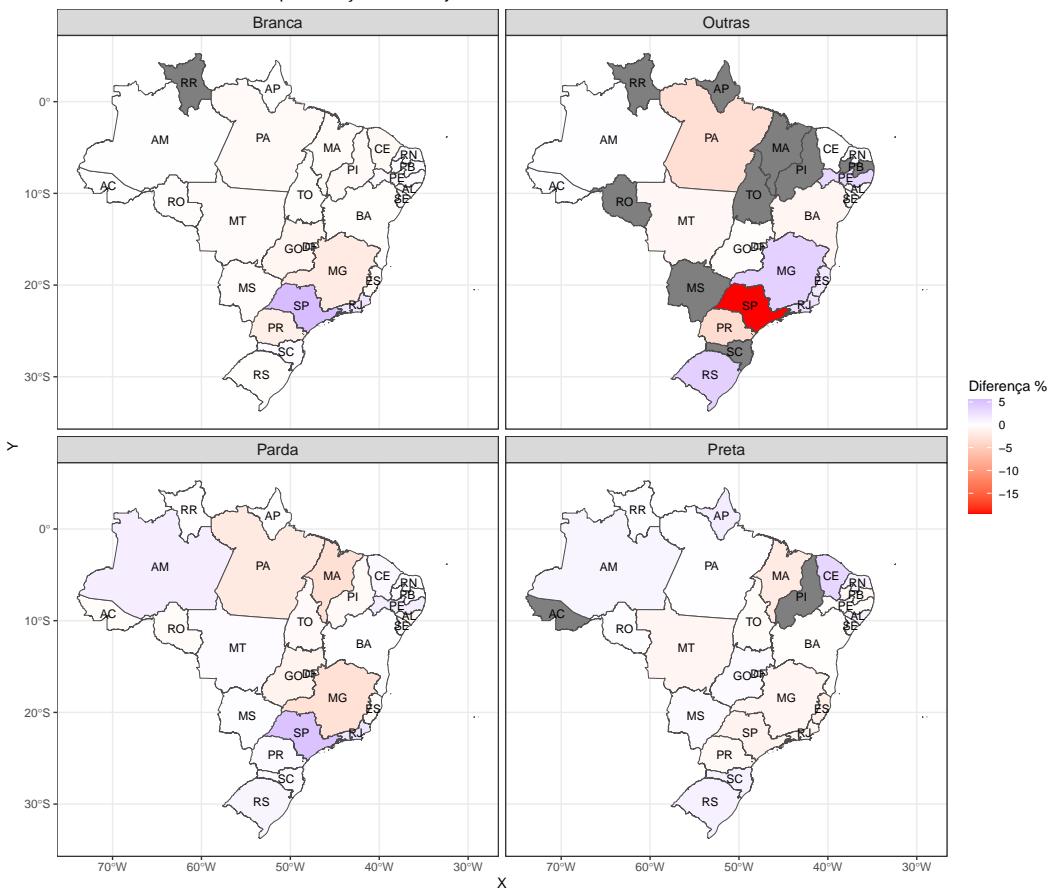
```
comparar_variavel(df, estratos, "cor_raca")
```

**Distribuição por cor\_raca : Survey vs. PNAD**



### Diferença percentual entre Survey e PNAD por UF e categoria de cor\_raca

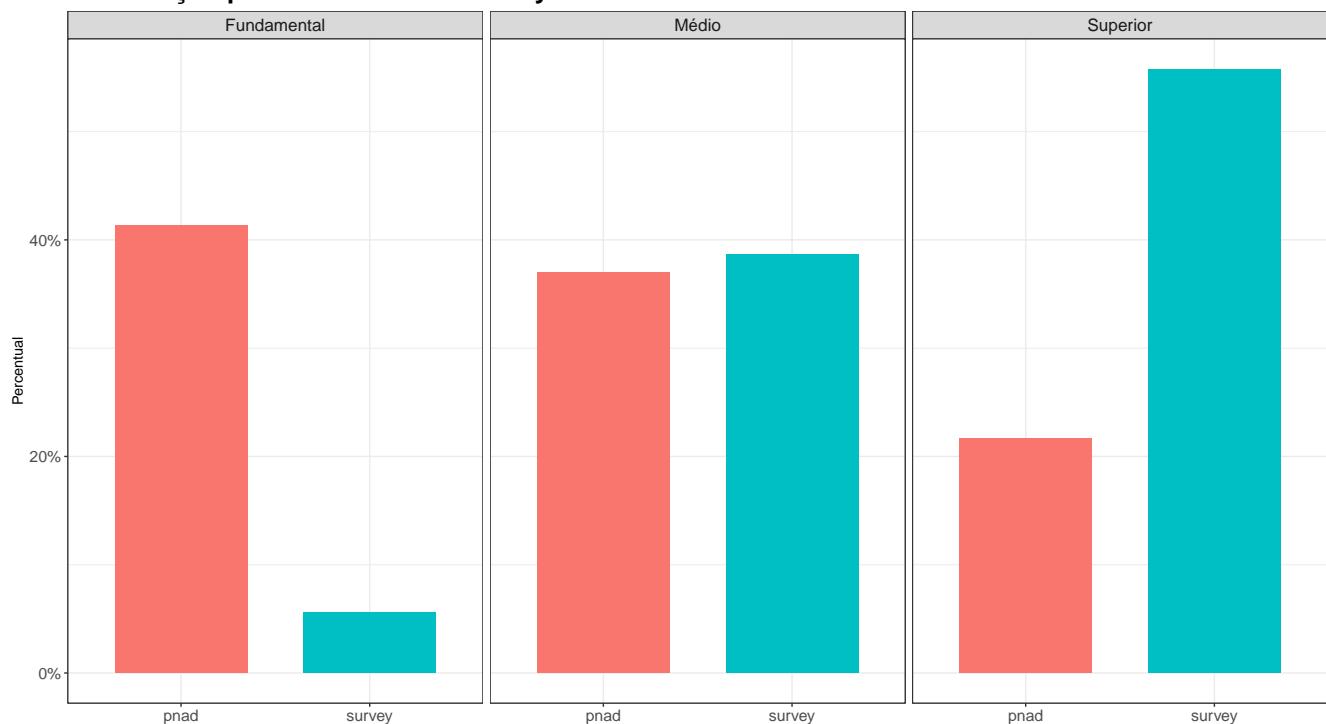
Valores > 0 indicam sobre-representação no survey



Por último, observa-se uma diferença crítica na escolaridade. Em particular, indivíduos com ensino superior são altamente sobre-representados em detrimento de indivíduos com ensino fundamental – uma característica bastante comum em amostras não probabilísticas extraídas *online*. Em vários estados da região Norte, inclusive, não houve registro de respondentes com ensino fundamental.

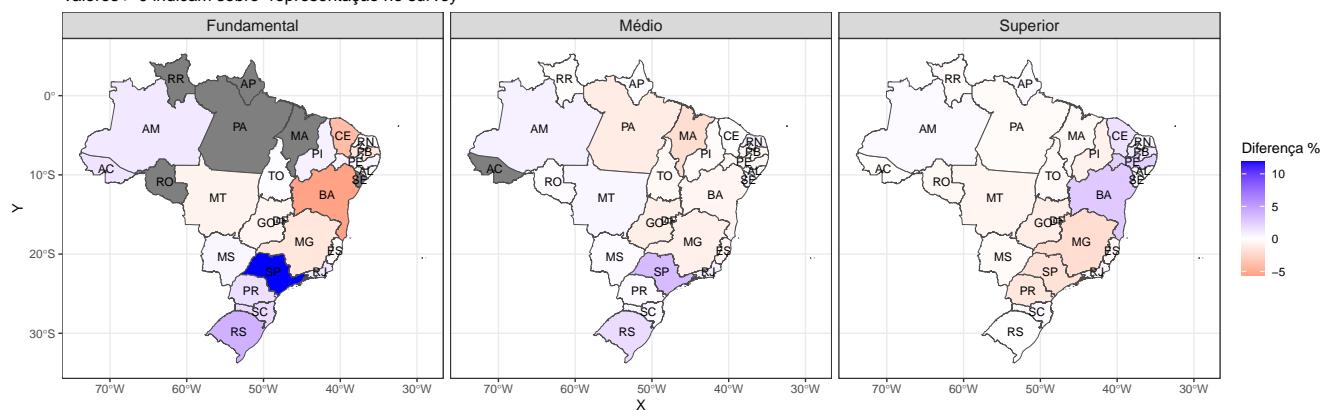
```
comparar_variavel(df, estratos, "escolaridade")
```

### Distribuição por escolaridade : Survey vs. PNAD



### Diferença percentual entre Survey e PNAD por UF e categoria de escolaridade

Valores > 0 indicam sobre-representação no survey



## Análise exploratória

Agora que já analisamos devidamente as diferenças entre a amostra coletada e a população brasileira, vamos passar a uma etapa crucial ao ajuste de modelos de regressão para pós-estratificação: a análise exploratória. O que acontece, na prática, é que a performance do MrP é altamente condicionada pela especificação do modelo. Isso significa, portanto – e conforme insistem Ghitza e Gelman (2013) –, que devemos analisar cuidadosamente quais variáveis incluiremos no modelo e, mais ainda, a maneira como essas variáveis se relacionam entre si.

A abordagem utilizada nesta tarefa foi a seguinte: entendendo que o voto em Bolsonaro em 2018 e o posicionamento contrário ao aborto são ações mais ou menos correlacionadas, implementei uma série de modelos multinível que tinham como objetivo inicial estimar a proporção da população que havia votado em Bolsonaro no primeiro turno das eleições presidenciais de 2018. Essa abordagem é útil porque, na prática, o percentual de indivíduos que votaram em Bolsonaro no primeiro turno das eleições de 2018 é um parâmetro conhecido: 46,03%. Após isso, utilizei o modelo que apresentou a melhor performance para estimar a proporção de indivíduos favoráveis ao aborto.

Essa abordagem, é claro, tem uma série de limitações. Uma delas diz respeito ao fato de que o voto em Bolsonaro e o posicionamento contrário ao aborto são atitudes apenas parcialmente correlacionadas. Isso significa dizer que, ao assumir uma mesma estratégia de MrP para estimar os dois parâmetros quase certamente levará a algum tipo de erro. Mesmo assim, assumo que essa abordagem é a melhor possível dadas as informações disponíveis.

Além disso, é importante admitir que as especificações de modelagem propostas nesta lista poderiam se valer, em larga escala, de uma análise exploratória mais detalhada. Mesmo assim, foi possível estimar de maneira bastante aceitável a proporção de indivíduos que votaram em Bolsonaro no primeiro turno de 2018. Espero, com isso, ter me aproximado de maneira mais ou menos razoável da taxa de indivíduos contrários ao aborto.

A função abaixo nos permite fazer alguns cruzamentos para observar quais variáveis estão correlacionadas ao voto em Bolsonaro. Isso será útil, especialmente, para especificar o modelo de regressão multinível.

```
cruzar_voto_bolsonaro <- function(df, var1, var2 = NULL) {  
  v1 <- sym(var1)  
  
  # ===== GRÁFICO DE BARRAS =====  
  if (is.null(var2)) {  
    dados <- df %>%  
      filter(!is.na(votou_bolsonato_1t_2018), !is.na (!!v1)) %>%  
      group_by (!!v1) %>%  
      summarise(  
        prop_bolsonaro = mean(votou_bolsonato_1t_2018),  
        .groups = "drop"  
    )  
  
    grafico <- ggplot(dados, aes(x = !!v1, y = prop_bolsonaro)) +  
      geom_col(fill = "steelblue") +  
      scale_y_continuous(labels = percent_format(accuracy = 1)) +  
      labs(  
        title = paste("Proporção de votos em Bolsonaro (1º turno - 2018) por", var1),  
        subtitle = "Cruzamento com",  
        x_label = "Posicionamento  
        contrário ao  
        aborto")  
  } else {  
    dados <- df %>%  
      filter(!is.na(votou_bolsonato_1t_2018), !is.na (!!v1), !is.na (!!var2)) %>%  
      group_by (!!v1, !!var2) %>%  
      summarise(  
        prop_bolsonaro = mean(votou_bolsonato_1t_2018),  
        .groups = "drop"  
    )  
  
    grafico <- ggplot(dados, aes(x = !!v1, y = prop_bolsonaro)) +  
      geom_col(fill = "steelblue") +  
      facet_grid(~ !!var2) +  
      scale_y_continuous(labels = percent_format(accuracy = 1)) +  
      labs(  
        title = paste("Proporção de votos em Bolsonaro (1º turno - 2018) por", var1),  
        subtitle = "Cruzamento com",  
        x_label = "Posicionamento  
        contrário ao  
        aborto")  
  }  
}
```

```

    x = var1,
    y = "Proporção"
) +
theme_bw() +
theme(
  plot.title = element_text(size = 16, face = "bold")
)
} else {
v2 <- sym(var2)

dados <- df %>%
  filter(!is.na(votou_bolsonato_1t_2018), !is.na (!!v1), !is.na (!!v2)) %>%
  group_by (!!v1, !!v2) %>%
  summarise(
    prop_bolsonaro = mean(votou_bolsonato_1t_2018),
    .groups = "drop"
  )

grafico <- ggplot(dados, aes(x = !!v2, y = prop_bolsonaro, fill = !!v2)) +
  geom_col(position = "dodge") +
  facet_wrap(as.formula(paste("~", var1))) +
  scale_y_continuous(labels = percent_format(accuracy = 1)) +
  labs(
    title = paste("Proporção de votos em Bolsonaro (1º turno - 2018) por", var1, "e", var2),
    x = var2,
    y = "Proporção"
) +
theme_bw() +
theme(
  legend.position = "none",
  strip.text = element_text(size = 12),
  plot.title = element_text(size = 16, face = "bold")
)
}

print(grafico)

# ===== MAPA FACETADO POR CATEGORIA DE var1 =====
votos_por_uf <- df %>%
  filter(!is.na(votou_bolsonato_1t_2018), !is.na (!!v1), !is.na(uf)) %>%
  group_by(uf, !!v1) %>%
  summarise(
    prop_bolsonaro = mean(votou_bolsonato_1t_2018),
    .groups = "drop"
  ) %>%
  mutate(categoria = as.character (!!v1))

```

```

mapa_df <- ufs_sf %>%
  left_join(votos_por_uf, by = c("abbrev_state" = "uf"))

# Centroides para siglas
centroides <- st_centroid(ufs_sf) %>%
  st_coordinates() %>%
  as_tibble() %>%
  bind_cols(ufs_sf["abbrev_state"])

mapa <- ggplot(mapa_df) +
  geom_sf(aes(fill = prop_bolsonaro), color = "gray30", size = 0.2) +
  geom_text(
    data = centroides,
    aes(X, Y, label = abbrev_state),
    size = 3, color = "black"
  ) +
  scale_fill_gradient(low = "white", high = "blue", labels = percent_format(accuracy = 1)) +
  facet_wrap(~ categoria) +
  labs(
    title = paste("Proporção de votos em Bolsonaro por UF e categoria de", vari),
    fill = "Proporção"
  ) +
  theme_bw() +
  theme(
    strip.text = element_text(size = 10),
    plot.title = element_text(size = 16, face = "bold")
  )

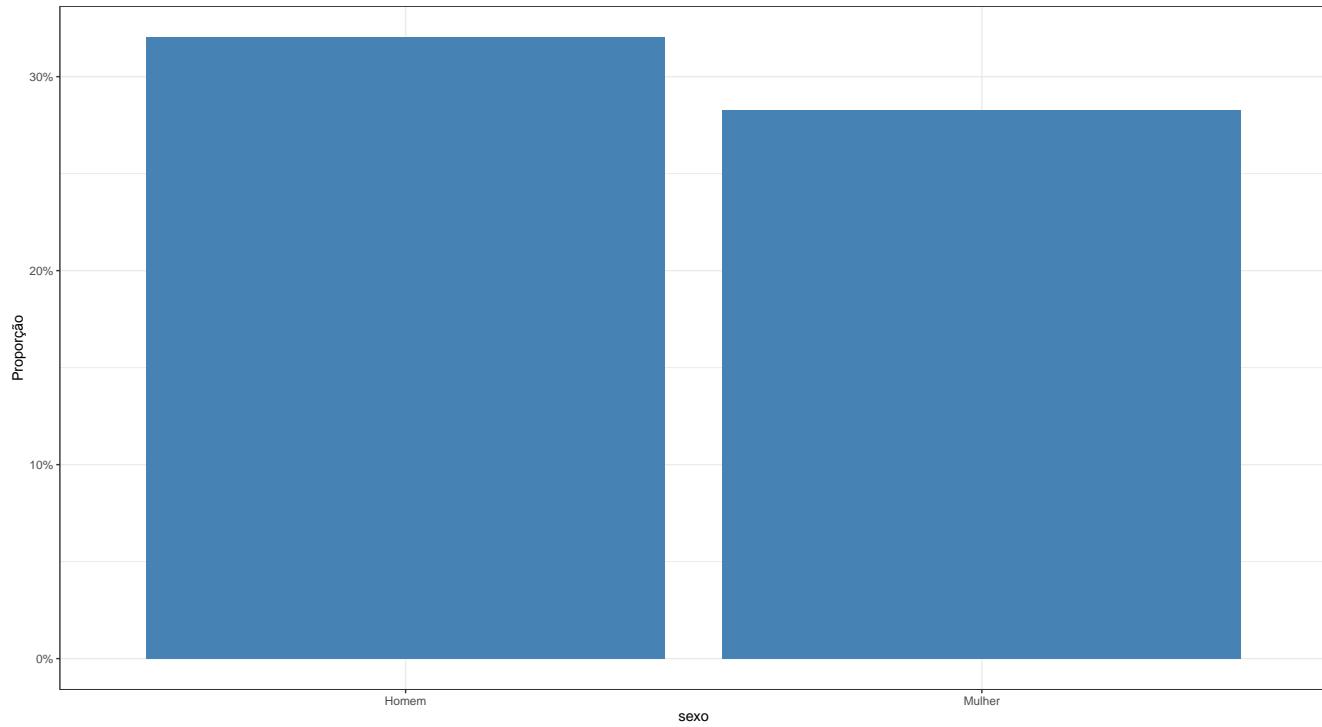
  print(mapa)
}

```

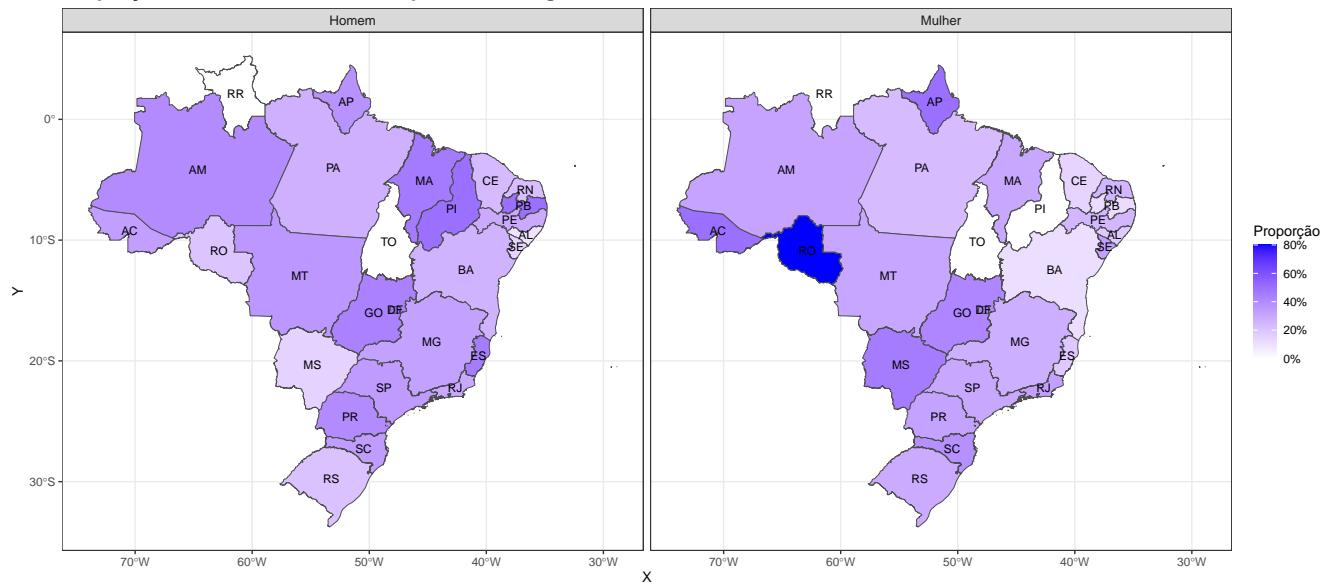
Abaixo, apresento a proporção de homens e mulheres, na amostra, que votaram em Bolsonaro, e a mesma informação por estado. Em geral, note que não há diferenças muito substantivas, embora a proporção geral de voto em Bolsonaro seja menor entre a população feminina.

```
cruzar_voto_bolsonaro(df, "sexo")
```

**Proporção de votos em Bolsonaro (1º turno – 2018) por sexo**

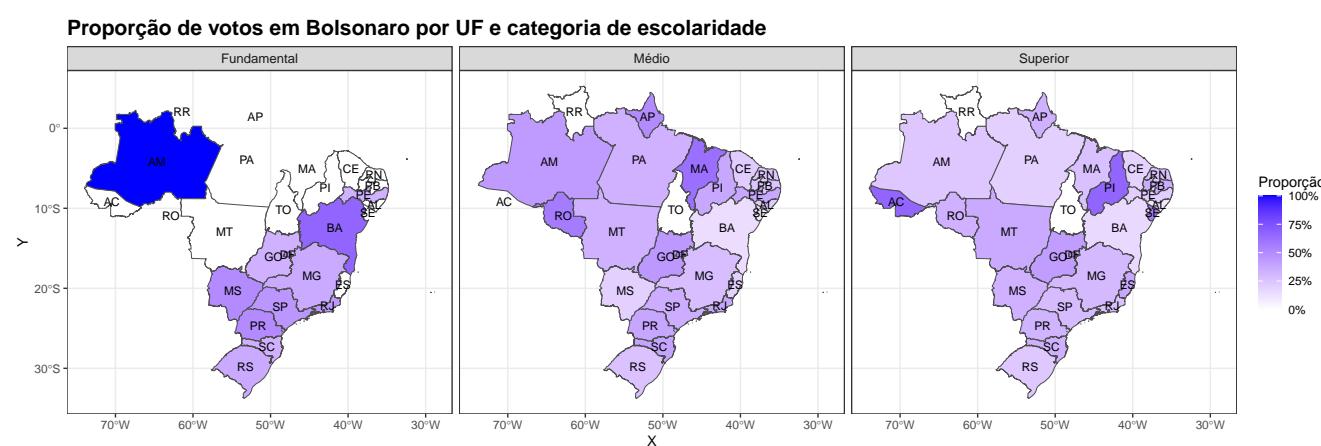
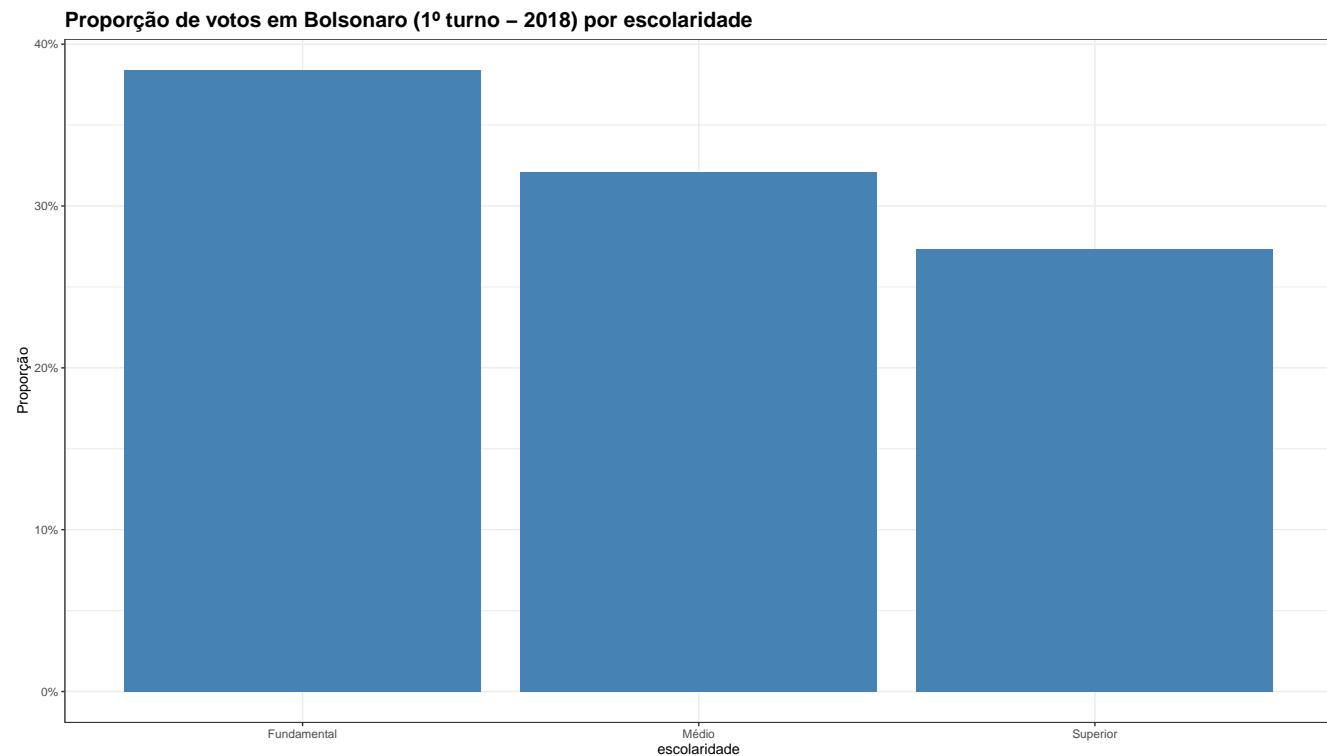


**Proporção de votos em Bolsonaro por UF e categoria de sexo**



Em termos de escolaridade, as diferenças são substantivas. Em especial, vale notar que indivíduos com ensino fundamental, na amostra coletada, votam proporcionalmente mais em Bolsonaro. É oportuno relembrar, neste ponto, que a proporção de indivíduos com ensino fundamental é altamente sub-representada na amostra em relação à população.

```
cruzar_voto_bolsonaro(df, "escolaridade")
```



De fato, como podemos observar na tabela abaixo, a nossa amostra é bem menos bolsonarista do que deveria. Apenas 30% dos respondentes afirmam ter votado em Bolsonaro no primeiro turno.

Distribuição do voto em Bolsonaro no 1º turno de 2018  
Estimativas 'puras' com base na amostra enviesada

Votou em Bolsonaro no 1º turno (2018)	Proporção
0	70.2%
1	29.8%

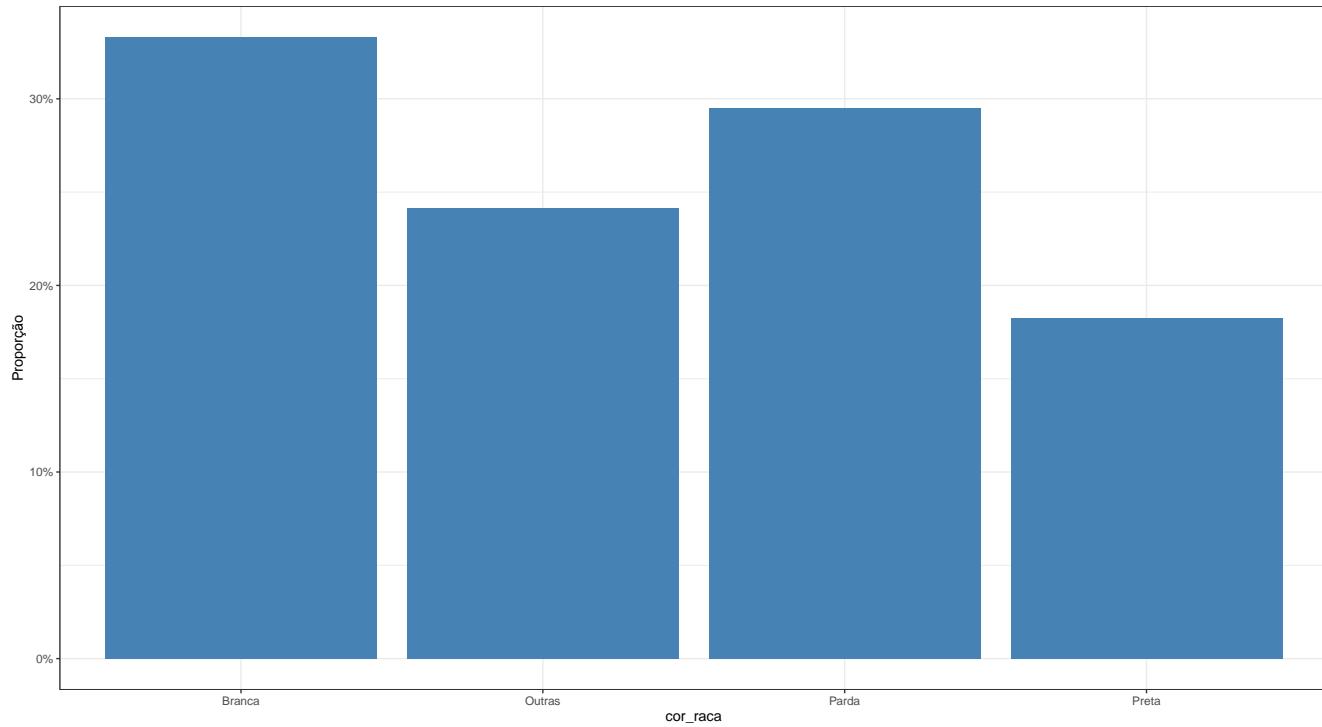
```
# Calcule a tabela de proporções
tabela_prop <- prop.table(table(df$votou_bolsonato_1t_2018)) %>%
  as.data.frame() %>%
  rename(Voto = Var1, Proporcao = Freq)

# Formate a tabela usando gt
tabela_prop %>%
  gt() %>%
  fmt_percent(columns = Proporcao, decimals = 1) %>%
  cols_label(
    Voto = "Votou em Bolsonaro no 1º turno (2018)",
    Proporcao = "Proporção"
  ) %>%
  tab_header(
    title = "Distribuição do voto em Bolsonaro no 1º turno de 2018",
    subtitle = "Estimativas 'puras' com base na amostra enviesada"
  )
```

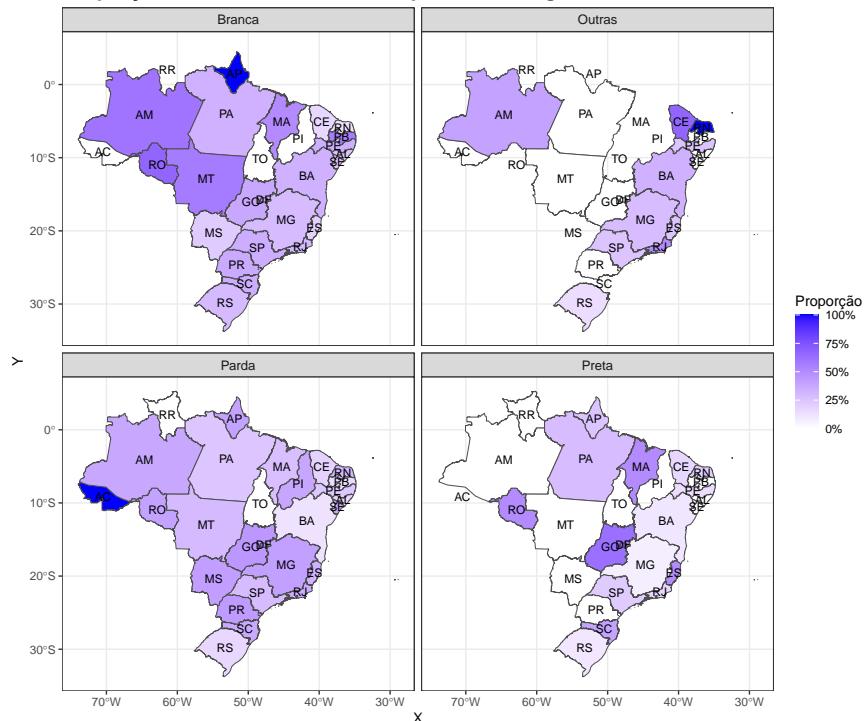
No que diz respeito à cor/raça na amostra, brancos são os que mais votam em Bolsonaro; pretos, por outro lado, são os que votam menos. Não há diferenças significativas por estados, embora brancos pareçam votar mais em Bolsonaro nos estados do Amazonas, Rondônias e Mato-Grosso. Esse efeito pode, na realidade, ser apenas fruto de amostra limitada.

```
cruzar_voto_bolsonaro(df, "cor_raca")
```

**Proporção de votos em Bolsonaro (1º turno – 2018) por cor\_raca**



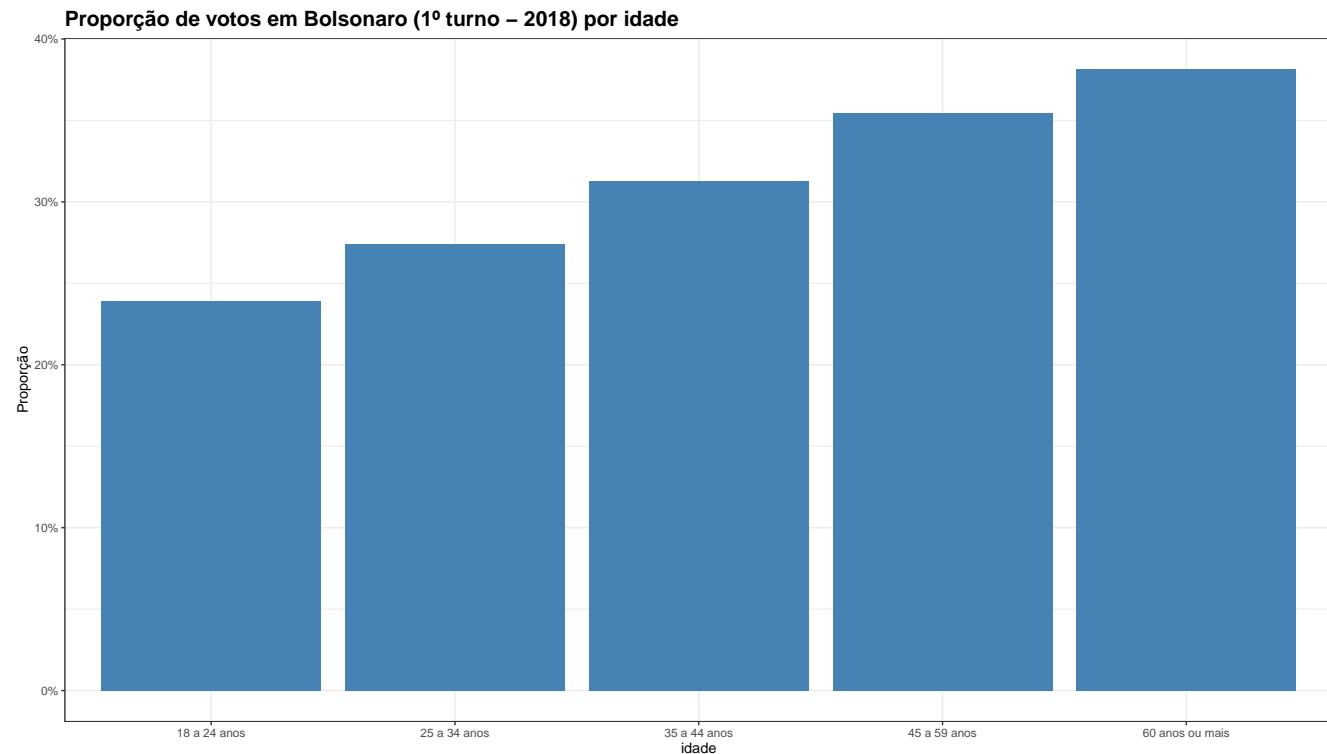
**Proporção de votos em Bolsonaro por UF e categoria de cor\_raca**



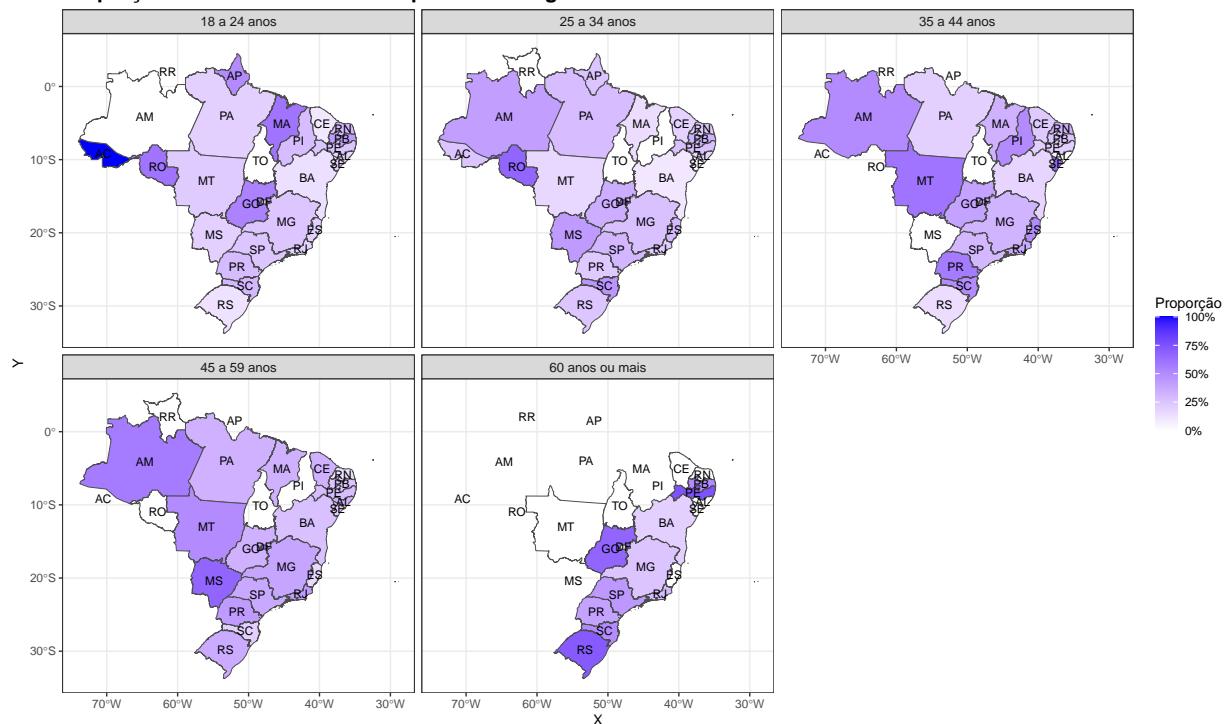
Em relação à idade, a relação é nítida: quanto maior a idade, maior é a proporção de votos em Bolsonaro. Quebrando por estado, parece haver uma certa variação especialmente nas faixas de 35 a 44 anos e 45 a 59 anos, que tendem a ser um pouco mais bolsonaristas em alguns estados da região Sul e da região Centro-Oeste. O mesmo vale para a amostra de indivíduos de 60 anos ou mais, mas não é possível inferir

muito de uma amostra tão pequena.

```
cruzar_voto_bolsonaro(df, "idade")
```



**Proporção de votos em Bolsonaro por UF e categoria de idade**



## Resultados

Considerando a análise exploratória anterior, propus duas abordagens de MrP. A primeira abordagem é bem simples e inclui apenas interceptos variáveis para cada variável de perfil sociodemográfico: estado, sexo, cor/raça, idade e escolaridade. Como é possível observar, o desempenho ao estimar a proporção de indivíduos que votaram em Bolsonaro é ruim, bem distante do valor verdadeiro do parâmetro:

```
model1 <- glmer(
  votou_bolsonato_1t_2018 ~
    (1 | uf) +
    (1 | sexo) +
    (1 | cor_raca) +
    (1 | idade) +
    (1 | escolaridade),
  data = df,
  family = binomial(link = "logit")
)

estratos$prob_voto_bolsonaro_1t_2018 <- predict(model1,
                                                 newdata = estratos,
                                                 type = "response",
                                                 re.form = NULL,
                                                 allow.new.levels = TRUE)

sum(estratos$prob_voto_bolsonaro_1t_2018 * estratos$n) / sum(estratos$n)
```

[1] 0.3336922

A segunda abordagem é um pouco mais complexa: além de incluir interceptos variáveis para algumas das covariáveis, incluo também *slopes* variáveis para outras – em particular, idade e escolaridade. De fato, a inclusão de uma *varying slope* para a idade foi justificada pela análise exploratória; a inclusão da *varying slope* para a escolaridade, por outro lado, foi resultado da testagem de uma série de especificações distintas.

```
model2 <- glmer(
  votou_bolsonato_1t_2018 ~
    (1 | sexo) +
    (1 | cor_raca) +
    (1 + idade | uf) +
    (1 + escolaridade | uf),
  data = df,
  family = binomial(link = "logit"),
  control = glmerControl(optimizer = "bobyqa")
)

estratos$prob_voto_bolsonaro_1t_2018 <- predict(model2,
```

```

            newdata = estratos,
            type = "response",
            re.form = NULL,
            allow.new.levels = TRUE)

print(sum(estratos$prob_voto_bolsonaro_1t_2018 * estratos$n) / sum(estratos$n))

```

[1] 0.4607143

Com o modelo acima, chegamos muito perto (excessivamente perto, provavelmente) da proporção de indivíduos que votaram em Jair Bolsonaro. Para todos os efeitos, utilizarei essa especificação para estimar a proporção de indivíduos que são contrários ao aborto:

```

model3 <- glmer(
  contra_aborto ~
    (1 | sexo) +
    (1 | cor_raca) +
    (1 + idade | uf) +
    (1 + escolaridade | uf),
  data = df,
  family = binomial(link = "logit"),
  control = glmerControl(optimizer = "bobyqa")
)

estratos$prob_contra_aborto <- predict(model3,
                                         newdata = estratos,
                                         type = "response",
                                         re.form = NULL,
                                         allow.new.levels = TRUE)

print(sum(estratos$prob_contra_aborto * estratos$n) / sum(estratos$n))

```

[1] 0.670788

Estimamos que a proporção da população que é contra o aborto é de aproximadamente 67%. Essa taxa é consistente com uma pesquisa realizada pelo Datafolha em 2024, conforme [este link](#), e está relativamente próxima de estimativas um pouco mais antigas, como a da Quaest, indicada [neste outro link](#).

Abaixo, apresento as estimativas por região do Brasil:

```

# Calcular proporção contra aborto por região
regiao_resultado <- estratos %>%
  group_by(regiao) %>%
  summarise(
    pct_contra_aborto = sum(prob_contra_aborto * n) / sum(n)
  ) %>%

```

População contrária à Legalização do Aborto por Região  
Estimativas ponderadas por MrP

Região	% Contra Aborto (Pós-Estratificado)
Centro-Oeste	74.9
Nordeste	65.0
Norte	51.4
Sudeste	66.3
Sul	77.5

```

mutate(
  pct_contra_aborto = pct_contra_aborto * 100 # converter para percentual
)

# Apresentar com gt
regiao_resultado %>%
  gt() %>%
  fmt_number(columns = pct_contra_aborto, decimals = 1) %>%
  cols_label(
    regiao = "Região",
    pct_contra_aborto = "% Contra Aborto (Pós-Estratificado)"
  ) %>%
  tab_header(
    title = "População contrária à Legalização do Aborto por Região",
    subtitle = "Estimativas ponderadas por MrP"
)

```

Dividindo os resultados por região, observamos que o posicionamento contrário ao aborto é mais alta nos estados do Sul e do Centro-Oeste, e mais baixa na região Norte.

## Considerações finais

De maneira geral, observamos que a pós-estratificação utilizando um modelo de MrP relativamente complexo, com *slopes* variáveis, permitiu estimar bem a proporção de indivíduos que votaram em Jair Bolsonaro. Aplicando este modelo para estimar a proporção da população que é contrária ao aborto, foi possível obter estimativas consistentes com as publicadas por alguns dos principais institutos de pesquisa do país. Esses resultados são um exemplo de como amostras não probabilísticas e pouco representativas podem ser trabalhadas para estimar, com alguma qualidade, parâmetros populacionais desconhecidos.

## Referências

Gelman, Andrew, e Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Ghitza, Yair, e Andrew Gelman. 2013. «Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups». *American Journal of Political Science* 57 (3): 762–76. <https://doi.org/10.1111/ajps.12004>.

Smith, Amy Erica, e Taylor C. Boas. 2024. «Religion, Sexuality Politics, and the Transformation of Latin American Electorates». *British Journal of Political Science* 54: 816–35. <https://doi.org/10.1017/S0007123423000613>.