Tarefa de survey

Felipe Lamarca

2025-06-05

Nesta tarefa implementamos 3 desenhos amostrais (AAS, AAS com estratificação e por conglomerados) usando dados de Macaé extraídos do Censo de 2010. O objetivo é avaliar o desempenho de cada desenho em vista da estimação da proporção de indivíduos alfabetizados no município. Observamos que a estimação no desenho estratificado é melhor que no desenho AAS do ponto de vista da variância da estimativa, mas há espaço para melhora a partir da escolha de variáveis de estratificação mais adequadas. O desenho por conglomerado, como esperado, possui variância amostral maior que o desenho AAS. A magnitude dessa diferença varia de acordo com diferentes escolhas de tamanho de conglomerados e número de entrevistas em cada conglomerado.

Introdução

Nesta tarefa, vamos implementar versões de três dos principais desenhos amostrais utilizados em pesquisas de survey, conforme visto em Groves et al. (2011). Para isso, partiremos da base `macae.rds`, proveniente do Censo de 2010, que consiste em um banco de dados de aproximadamente 200 mil linhas, em que cada linha representa um habitante do município de Macaé, no estado do Rio. Temos uma série de informações importantes sobre cada habitante: o local de moradia (nos níveis de distrito, subdistrito, bairro e setor censitário) e uma variável binária `alfabetizado`, que indica se o indivíduo é (1) ou não (0) alfabetizado.

Nosso objetivo é, utilizando diferentes desenhos amostrais, estimar tão bem quanto possível a taxa de alfabetização da população do município. Como estamos utilizando os dados de um censo, não esperamos lidar com erros sistemáticos de cobertura – i.e., assumimos que temos um bom quadro amostral, que inclui toda a população do município. De fato, tratando-se dos dados de um censo demográfico, já conhecemos a priori a estatística "verdadeira", ou seja, a taxa de alfabetização da população de Macaé. Mesmo assim, a implementação de alguns desenhos amostrais será útil para nos permitir identificar abordagens mais ou menos eficientes (neste caso, com maior ou menor variância) em vista da estimação do parâmetro populacional.

Iniciamos a implementação dos desenhos amostrais com uma amostra aleatória simples. Como veremos, esse é o tipo de amostragem mais simples e permite a derivação de estatísticas amostrais de maneira muito direta; por outro lado, tendemos a obter alta variância na medida em que estamos sujeitos a selecionar grupos muito diferentes entre si e, mais do que isso, temos pouco controle sobre a representação dos grupos em cada realização da amostra.

Com isso em mente, tentaremos visualizar o efeito positivo de implementar uma amostra aleatória simples com estratificação utilizando alocação proporcional. Neste caso, procuramos uma variável de estratificação

que minimiza a variância da variável de interesse dentro dos estratos e extraímos amostras aleatórias simples dentro desses estratos; portanto, esperamos obter uma menor variância do que na amostra aleatória simples (Groves et al. 2011, 119).

Por fim, implementamos um desenho amostral por conglomerados, em que primeiro amostramos os conglomerados com probabilidade de seleção proporcional ao tamanho da população e, em seguida, amostramos indivíduos dentro desses conglomerados. Embora essa seja uma abordagem amplamente utilizada por acarretar sensível diminuição de custos, trata-se de um desenho conhecido por aumentar a variância em relação à amostra aleatória simples (Groves et al. 2011, 112–13).

Na ocasião da implementação de cada um dos desenhos amostrais, discutiremos com mais detalhes alguns pontos teóricos e práticos importantes. Finalizamos a discussão com uma comparação das distribuições amostrais da estatística de interesse em cada um dos desenhos, comparando nossos resultados às expectativas sugeridas pela literatura.

Implementação dos desenhos amostrais

Desenho AAS

Em um desenho de amostra aleatória simples, cada elemento da população tem a mesma probabilidade de ser incluído na amostra. Mais especificamente, a probabilidade de inclusão é dada por:

$$\pi_i = \frac{n}{N},$$

onde n é o número de elementos na amostra e N é o número de elementos na população. A probabilidade de uma particular amostra ser selecionada é dada por:

$$\mathbb{P}(S) = \frac{1}{\binom{N}{n}},$$

sendo que $\binom{N}{n} = \frac{N!}{n! \times (N-n)!}$.

Trata-se de uma amostragem de natureza totalmente aleatória e que garante, para N grande, que a probabilidade de se obter uma particular amostra seja muito pequena. No caso dos dados de Macaé, vemos que a chance de obtermos uma amostra qualquer de tamanho $n = 1200$ a partir do nosso banco de dados é baixíssima:

```
prob_S <- 1 / choose(nrow(macae), 1200)
cat(sprintf("Probabilidade de selecionar uma particular amostra: %.30f\n", prob_S))
```

```
Probabilidade de selecionar uma particular amostra: 0.000000000000000000000000000000
```

Vamos simular 2000 amostras diferentes (i.e., sem reposição) com $n = 1200$ e guardar a média de cada uma das realizações:

```
alfabetizada_srs <- numeric(2000)

for (i in 1:2000) {
  sample <- macae %>%
    slice_sample(n=1200, replace=FALSE)

  alfabetizada_srs[[i]] <- mean(sample$alfabetizada)
}

var_srs <- var(alfabetizada_srs)
```

Extraídas as amostras, calculamos também a variância da estatística de interesse, que, no caso do desenho AAS, servirá de benchmark no cálculo do *design effect* dos demais desenhos a seguir.

Desenho AAS com estratificação

Extrair amostras aleatórias simples com estratificação consiste em dividir os dados em grupos (estratos) e extrair amostras aleatórias simples dos indivíduos dentro desses grupos. Há algumas técnicas possíveis utilizadas para alocar o tamanho n_h de indivíduos a serem amostrados em cada estrato h, sendo uma das preferidas a alocação proporcional. Nesse caso, $n_h = n \times \frac{N_h}{N}$, sendo n o número de elementos na amostra, N_h a população do estrato h e N a população total.

No fim das contas, a consequência dessa abordagem é que cada estrato é representado na amostra exatamente em proporção ao seu tamanho na população. Mais ainda, se esses estratos forem "substantivamente significativos" (Groves et al. 2011, 120) – ou seja, difiram entre si ao mesmo tempo em que se assemelham internamente no que diz respeito à variável de interesse –, podemos esperar uma variância menor do que no desenho que usa puramente a amostragem aleatória simples. Trata-se de um resultado lógico na medida em que, ao dividir o espaço amostral em grupos relativamente homogêneos internamente, a variância dentro dos grupos é menor; consequentemente, a variância amostral (resultante da combinação das variâncias dos grupos) também será menor (Groves et al. 2011, 119).

Pelos motivos discutidos a seguir, optamos por agrupar os dados em subdistritos (e distritos, consequentemente):

```
# agrupamos os dados em distritos e subdistritos
n_entrevistas <- macae %>%
  group_by(distrito, subdistrito) %>%
  summarise(
    Nh = n(),
    prop_alfabetizada = mean(alfabetizada),
    var_alfabetizada = var(alfabetizada),
    .groups = "drop"
  ) %>%
  mutate(
    # alocacao proporcional
```

```
    Wh = Nh / sum(Nh),
    nh = round(1200 * Wh)
  )

# trazemos o numero de entrevistas para o banco original
macae_estratificada <- macae %>%
  inner_join(n_entrevistas, by=c("distrito", "subdistrito"))

# instanciamos a lista que guardara as medias
alfabetizada_estratificada <- numeric(2000)

# amostramos 2000 vezes
# group map retorna uma lista de tibbles (cada entrada sendo um tibble
# para cada categoria no agrupamento), que juntamos com bind_rows --
# i.e., uma amostra de tamanho n = nh para cada subdistrito
for (i in 1:2000) {
  sample <- macae_estratificada  %>%
    group_by(distrito, subdistrito) %>%
    group_map(
      ~ slice_sample(.x, n = unique(.x$nh), replace=FALSE),
      .keep = TRUE
    ) %>%
    bind_rows()

  alfabetizada_estratificada[[i]] <- mean(sample$alfabetizada)
}

# calculamos a variancia da estatistica de interesse e o design effect
var_alfabetizada_estratificada <- var(alfabetizada_estratificada)
design_effect_alfabetizada_estratificada <- var_alfabetizada_estratificada / var_srs
cat(sprintf("Design effect -- distrito e subdistrito: %.3f\n", design_effect_alfabetizada_estrati
```

```
Design effect -- distrito e subdistrito: 0.930
```

Note que o *design effect* indica um resultado apenas marginalmente melhor em relação ao desenho AAS. De fato, sabemos que a escolha da variável de estratificação é fundamental na qualidade do resultado – escolhas ineficazes levam a resultados no máximo um pouco melhores em termos de variância. Para obter melhores resultados seria importante obter, por exemplo, um número menor de estratos (Groves et al. 2011, 118) e, idealmente, estratos com variância interna pequena. Vejamos a tabela com o número de entrevistas, que traz algumas estatísticas dos grupos:

```
print(n_entrevistas)
```

```
# A tibble: 11 x 7
   distrito   subdistrito    Nh prop_alfabetizada var_alfabetizada      Wh    nh
   <chr>      <chr>       <int>             <dbl>            <dbl>   <dbl> <dbl>
```

1	Cachoeiro~	Sem inform~	1319	0.765	0.180	0.00638	8
2	Córrego D~	Sem inform~	3992	0.789	0.166	0.0193	23
3	Frade	Sem inform~	1390	0.849	0.128	0.00672	8
4	Glicério	Sem inform~	2797	0.837	0.137	0.0135	16
5	Macaé	Aeroporto	37760	0.880	0.106	0.183	219
6	Macaé	Barra De M~	32362	0.839	0.135	0.157	188
7	Macaé	Cabiúnas	23952	0.842	0.133	0.116	139
8	Macaé	Centro	79381	0.893	0.0957	0.384	461
9	Macaé	Imboassica	20132	0.901	0.0892	0.0974	117
10	Macaé	Nova Cidade	2095	0.830	0.141	0.0101	12
11	Sana	Sem inform~	1548	0.846	0.131	0.00749	9

Observe que há grupos com poucas entrevistas alocadas, sendo a maioria deles na categoria "Sem informação" nos subdistritos. Além disso, o Centro de Macaé concentra boa parte das entrevistas. A despeito de uma série de tentativas de criação de novos agrupamentos com o objetivo de minimizar a variância interna, diminuir o número de estratos e homogeneizar o tamanho dos grupos, não tive sucesso em encontrar um *design effect* inferior ao reportado no desenho acima. Para citar alguns exemplos, tentativas incluem a divisão dos bairros em quartis de acordo com a proporção de alfabetização, o agrupamento dos subdistritos "Sem informação", e outros agrupamentos de acordo com a proporção de alfabetização.

Vale notar, ainda, que o uso da variável `bairro` no processo de estratificação levava a um *design effect* menor. No entanto, como a amostra é composta por mais de 30 bairros, essa escolha poderia dificultar a operacionalização da pesquisa – inclusive pelo fato de que teríamos alguns bairros com pouquíssimas entrevistas alocadas.

Desenho por conglomerados

A amostragem por conglomerados é normalmente implementada em duas etapas: na primeira amostramos o número n de conglomerados (no nosso caso, setores censitários) que irão compor a amostra e, depois, amostramos dentro desses setores um número n_i fixo de pessoas a serem amostradas. Quando o número de elementos dentro de cada conglomerado não é fixo, é interessante definir a probabilidade de inclusão do conglomerado na amostra de maneira proporcional ao seu tamanho – uma técnica chamada Probabilities Proportional to Size (PPS) (Groves et al. 2011, 126).

Devemos escolher o número de conglomerados em função do número n_i fixo de pessoas a serem amostradas em cada conglomerado, e o enunciado nos diz que devemos escolher entre 10 e 20 pessoas. Para obter uma amostra de 1200 indivíduos, os casos extremos envolveriam amostrar 120 setores censitários e amostrar 10 pessoas em cada, ou então amostrar 60 setores censitários e amostrar 20 pessoas em cada.

Essas opções implicam discutir o que leva alguém a escolher um desenho amostral por conglomerado e os prós e contras dessa escolha. Amostrar conglomerados ao invés de indivíduos facilita muito o processo de coleta de dados, por exemplo, ao delimitar as entrevistas a um número bem menor de setores censitários, o que diminui custos e facilita a logística. No entanto, devemos levar em conta o fato de que, provavelmente, integrantes de um mesmo conglomerado são mais ou menos parecidos, ao mesmo tempo que são razoavelmente diferentes dos integrantes dos demais conglomerados. Portanto, na medida em que escolhemos amostrar menos setores censitários e fazer mais entrevistas em cada um deles, a tendência é que a variância do desenho amostral aumente: estamos, ao mesmo tempo, (i) incluindo cada vez menos

informações "novas" na amostra ao fazer várias entrevistas em um mesmo setor, e (ii) deixando de coletar informações diferentes de outros setores (Groves et al. 2011, 112–13).

A escolha, portanto, deve ponderar custos operacionais e o efeito negativo que a escolha de um número pequeno de conglomerados tem sobre a variância do desenho. Nesse caso, escolhemos amostrar 80 setores censitários e amostrar 15 pessoas em cada um (de fato, uma média das opções disponíveis). Vamos começar calculando a probabilidade de inclusão de cada setor censitário na amostra de maneira proporcional ao seu tamanho:

```
# calculamos a probabilidade de seleção na amostra usando PPS
probabilidades <- macae %>%
  group_by(codigo_setor) %>%
  summarise(
    Nh = n()
  ) %>%
  mutate(
    prob_inclusao = Nh / (sum(Nh))
  )

head(probabilidades)
```

codigo_setor	Nh	prob_inclusao
330240305010001	588	0.0028443
330240305010002	624	0.0030185
330240305010003	431	0.0020849
330240305010004	248	0.0011996
330240305010005	438	0.0021187
330240305010006	814	0.0039375

Agora vamos extrair as amostras:

```
# instanciamos a lista que guarda as medias
alfabetizada_conglomerados_80 <- numeric(2000)

# amostramos 2000 vezes
for (i in 1:2000) {
  setores_censitarios <- probabilidades %>%
    slice_sample(n = 80, weight_by = prob_inclusao)

  sample <- macae %>%
    filter (codigo_setor %in% unique(setores_censitarios$codigo_setor)) %>%
    group_by(codigo_setor) %>%
    group_map(
      ~ slice_sample(.x, n = 15, replace = FALSE),
      .keep = TRUE
    ) %>%
```

```
    bind_rows()

  alfabetizada_conglomerados_80[[i]] <- mean(sample$alfabetizada)
}

var_conglomerados_80 <- var(alfabetizada_conglomerados_80)
design_effect_conglomerados_80 <- var_conglomerados_80 / var_srs
cat(sprintf("Design effect com 80 conglomerados (15 entrevistas em cada): %.3f\n", design_effect_
```

```
Design effect com 80 conglomerados (15 entrevistas em cada): 1.219
```

Apenas por curiosidade científica, podemos também calcular o *design effect* para os dois desenhos mais extremos, isto é, um (1) em que amostramos 120 setores censitários e fazemos 10 entrevistas em cada, e outro (2) em que amostramos 60 setores censitários e fazemos 20 entrevistas em cada. De acordo com a literatura, esperamos um *design effect* menor no primeiro desenho em relação ao segundo, embora ambos devam ter variância amostral maior que o desenho AAS (Groves et al. 2011, 112–13).

```
##### 120 conglomerados e 10 entrevistas em cada
alfabetizada_conglomerados_120 <- numeric(2000)

for (i in 1:2000) {
  setores_censitarios <- probabilidades %>%
    slice_sample(n = 120, weight_by = prob_inclusao)

  sample <- macae %>%
    filter (codigo_setor %in% unique(setores_censitarios$codigo_setor)) %>%
    group_by(codigo_setor) %>%
    group_map(
      ~ slice_sample(.x, n = 10, replace = FALSE),
      .keep = TRUE
    ) %>%
    bind_rows()

  alfabetizada_conglomerados_120[[i]] <- mean(sample$alfabetizada)
}

var_conglomerados_120 <- var(alfabetizada_conglomerados_120)
design_effect_conglomerados_120 <- var_conglomerados_120 / var_srs
cat(sprintf("Design effect com 120 conglomerados (10 entrevistas em cada): %.3f\n", design_effect
```

```
Design effect com 120 conglomerados (10 entrevistas em cada): 1.059
```

```
##### 60 conglomerados e 20 entrevistas em cada
alfabetizada_conglomerados_60 <- numeric(2000)
```

```r
for (i in 1:2000) {
  setores_censitarios <- probabilidades %>%
    slice_sample(n = 60, weight_by = prob_inclusao)

  sample <- macae %>%
    filter (codigo_setor %in% unique(setores_censitarios$codigo_setor)) %>%
    group_by(codigo_setor) %>%
    group_map(
      ~ slice_sample(.x, n = 20, replace = FALSE),
      .keep = TRUE
    ) %>%
    bind_rows()

  alfabetizada_conglomerados_60[[i]] <- mean(sample$alfabetizada)
}

var_conglomerados_60 <- var(alfabetizada_conglomerados_60)
design_effect_conglomerados_60 <- var_conglomerados_60 / var_srs
cat(sprintf("Design effect com 60 conglomerados (20 entrevistas em cada): %.3f\n", design_effect_
```

```
Design effect com 60 conglomerados (20 entrevistas em cada): 1.247
```

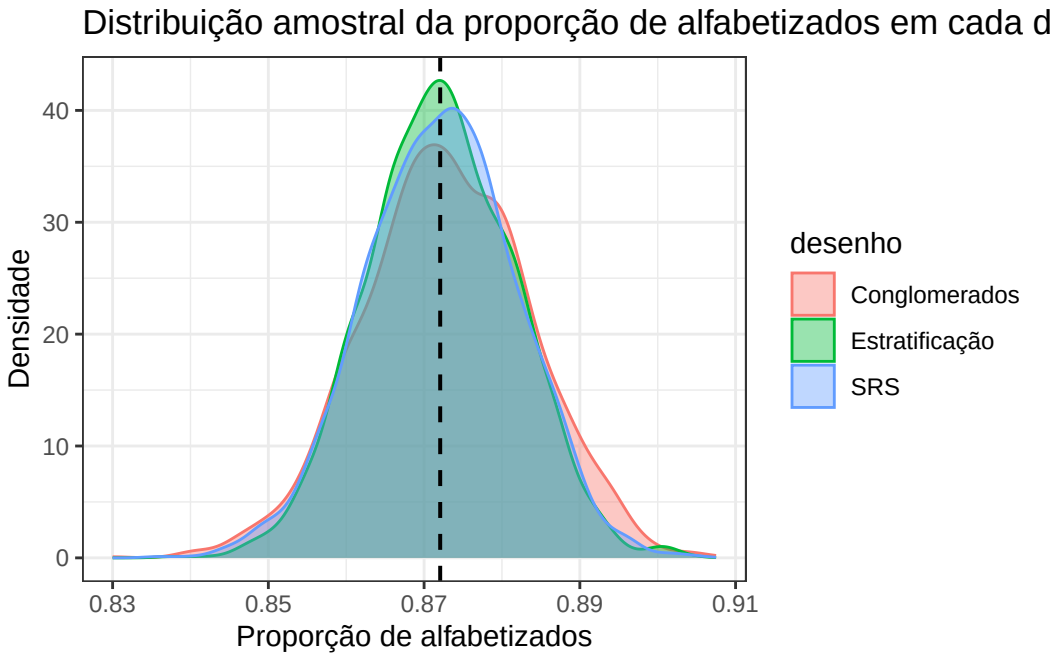
Resultados

Finalmente, com as amostras extraídas, vamos plotar a distribuição amostral da média de cada um dos desenhos:

```r
df_resultados <- tibble(
  valor = c(
    alfabetizada_srs,
    alfabetizada_estratificada,
    alfabetizada_conglomerados_80
    ),
  desenho = rep(c(
    "SRS",
    "Estratificação",
    "Conglomerados"
    ),
    each = 2000)
)

# Criar o gráfico com a linha vertical da média populacional
ggplot(df_resultados, aes(x = valor, fill = desenho, color = desenho)) +
  geom_density(alpha = 0.4) +
  geom_vline(xintercept = mean(macae$alfabetizada),
```

```
            linetype = "dashed",
            color = "black",
            size = 0.7) +
  labs(
    title = "Distribuição amostral da proporção de alfabetizados em cada desenho",
    x = "Proporção de alfabetizados",
    y = "Densidade"
  ) +
  theme_bw()
```



De fato, observamos que o desenho com estratificação por distritos e subdistritos possui uma variância amostral menor que o desenho AAS, com amostras mais frequentemente centradas em torno do parâmetro populacional verdadeiro. O desenho por conglomerados com 80 setores censitários e 15 entrevistas em cada possui, por outro lado, uma variância amostral maior – a distribuição da média amostral apresenta caudas mais pesadas, estimando a média populacional em valores ligeiramente mais extremos do que no desenho AAS.

Por fim, a Tabela 2 apresenta os *design effects*:

```
# Criar um data frame com os dados
design_effects_df <- data.frame(
  design = c(
    "Estratificação por distrito e subdistrito",
    "80 conglomerados (15 entrevistas em cada)",
    "120 conglomerados (10 entrevistas em cada)",
    "60 conglomerados (20 entrevistas em cada)"
  ),
  design_effect = c(
```

```
    design_effect_alfabetizada_estratificada,
    design_effect_conglomerados_80,
    design_effect_conglomerados_120,
    design_effect_conglomerados_60
  )
)

# Criar a tabela com o pacote gt
design_effects_df %>%
  arrange(design_effect) %>%
  gt() %>%
  cols_label(
    design = "Desenho amostral",
    design_effect = "Design effect"
  ) %>%
  tab_header(
    title = "Design effects por desenho amostral"
  ) %>%
  tab_options(
    table.font.size = "small",
    row.striping.include_table_body = TRUE
  )
```

Tabela 2: Design effects dos desenhos amostrais

Design effects por desenho amostral

Desenho amostral	Design effect
Estratificação por distrito e subdistrito	0.9304609
120 conglomerados (10 entrevistas em cada)	1.0592595
80 conglomerados (15 entrevistas em cada)	1.2192147
60 conglomerados (20 entrevistas em cada)	1.2472917

Em linha com o que foi observado no gráfico de distribuições e o que já foi discutido em ocasião anterior, obtivemos um design effect menor que 1 no desenho estratificado – mais especificamente, a variância amostral da média no desenho estratificado representa uma fração (grande, infelizmente) da variância do desenho AAS. Trata-se de um resultado marginalmente melhor, que poderia ser mais substantivo a partir da escolha de uma variável de estratificação mais adequada.

Além disso, encontramos *design effects* maiores que 1 em todos os desenhos por conglomerado, indicando que a variância amostral da média é maior nesses casos do que no desenho AAS. De fato, a discussão feita até aqui justifica esse resultado, afinal, a natureza desse desenho amostral normalmente implica maior variância. Observamos também que, quanto maior o número de conglomerados amostrados, menor é o *design effect*, embora todas as variâncias sejam maiores em relação ao desenho AAS. Trata-se de um resultado esperado na medida em que quanto mais informações novas (i.e., de conglomerados diferentes) somos capazes de inserir na amostra, menor é a variância esperada.

Referências

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, e Roger Tourangeau. 2011. *Survey methodology*. John Wiley & Sons.