Capítulo 7 Estimação para Domínios de Estudo

7.1 Domínios de estudo

Grande parte das pesquisas amostrais produz e divulga estimativas para certos subgrupos nos quais a população pode ser dividida. Tais subgrupos, aqui denominados *domínios* de estudo, *subpopulações* ou *pequenas áreas*, são quaisquer *subconjuntos da população* U para os quais desejamos obter ou produzir estimativas separadas a partir da amostra selecionada e observada dessa população.

Por exemplo, pesquisas domiciliares costumam apresentar estimativas para grupos tais como: homens de 40 anos ou mais, mulheres em idade reprodutiva (15 a 49 anos), crianças e adolescentes, etc. Pesquisas de empresas costumam apresentar estimativas por faixas de tamanho das empresas (por exemplo definido como o número de pessoas ocupadas nas empresas) ou classes de atividade econômica exercida. Em pesquisas agropecuárias, costuma-se apresentar estimativas por faixas de área total dos estabelecimentos (em hectares), ou segundo ocorrência de produção de certos produtos (por exemplo, para estabelecimentos com produção de café). Em todas estas, também é comum a apresentação de estimativas por regiões, unidades da federação, ou outras particições geográficas de interesse. Cada um dos subgrupos assim definidos configura um *domínio* de interesse.

Neste capítulo apresentamos as ideias centrais de como fazer *estimação para domínios*. Apresentaremos com detalhes o caso da estimação sob *Amostragem Aleatória Simples*, mas as ideias sobre como proceder com outros planos amostrais são as mesmas. A principal estratégia para estimar parâmetros em domínios de interesse é definir duas variáveis derivadas que podem ser calculadas para as unidades da amostra selecionada.

A primeira destas é simplesmente uma variável indicadora de pertinência ao domínio de interesse: seja d_i a variável indicadora do domínio d, isto é:

$$d_i = I(i \in U_d) = \begin{cases} 1 & \text{se } \ i \text{ pertence ao domínio } d, \\ 0 & \text{caso contrário.} \end{cases}$$

onde $U_d \subset U$ denota a parte da população U que forma o domínio d.

Denotamos por $N_d = \sum_{i \in U} d_i$ o tamanho do domínio d.

A segunda variável derivada é formada simplesmente pelo produto de duas outras variáveis de pesquisa: y_i e d_i, a saber:

$$y_{id} = y_i d_i = \begin{cases} y_i & \text{se } i \in U_d, \\ 0 & \text{caso contrário} \end{cases}$$

onde y é a variável de estudo cujos parâmetros se quer estimar para o domínio de interesse.

7.2 Parâmetros de interesse para domínios

Os parâmetros populacionais que usualmente se deseja estimar para um domínio genérico d são descritos a seguir.

A *proporção* de unidades populacionais no domínio d dada por:

$$p_d = \sum_{i \in U} d_i / N = N_d / N = \overline{D} \tag{7.1}$$

onde $\overline{D}$ é a média populacional da variável derivada d_i.

O *total* da variável y no domínio d dado por:

$$Y_d = \sum_{i \in U} y_{id} = \sum_{i \in U} y_i d_i = \sum_{i \in U_d} y_i \tag{7.2}$$

A *média* da variável y no domínio d dada por:

$$\overline{Y}_d = Y_d / N_d = \sum_{i \in U} y_{id} / \sum_{i \in U} d_i \tag{7.3}$$

Note que $\overline{Y}_d$ é um caso especial de *Razão de Médias* das variáveis y_{id} e d_i.

A *variância* da variável y no domínio d dada por:

$$S_d^2 = \sum_{i \in U} d_i (y_i - \overline{Y}_d)^2 / (N_d - 1) \tag{7.4}$$

Muitos outros parâmetros podem ser definidos para domínios, mas não é comum que os livros de Amostragem dediquem espaço ao tratamento de outros casos que não os considerados aqui. Isto se dá porque os casos de parâmetros que podem ser escritos como funções de totais populacionais são facilmente resolvidos empregando-se as ideias de estimação discutidas na seção 6.5.

7.3 Estimação de parâmetros para domínios sob AAS

Esta seção mostra como fazer para estimar os parâmetros populacionais para o domínio d considerados na seção 7.2, com base numa amostra aleatória simples sem reposição de tamanho n extraída da população U. Embora a discussão seja restrita aqui ao caso de amostras aleatórias simples, o processo de adaptação de estimadores para parâmetros de domínios aqui mostrado pode ser facilmente seguido para o caso de outros planos amostrais. Tudo se baseia na ideia de criação das variáveis derivadas apresentada na seção 7.1.

Passo 1: Selecionar uma AAS de tamanho n da população U de tamanho N, e observar (y_i) para todo $i \in s$.

Passo 2: Então construir as variáveis derivadas $d_i = I(i \in U_d)$ e $y_{id} = y_i d_i$.

Passo 3: Estimar os parâmetros de interesse, adaptando os estimadores apresentados nos capítulos 4, 5 e 6 para considerar as variáveis derivadas definidas.

Começamos tratando da estimação da proporção p_d de unidades no domínio, por ser o caso mais simples. Considerando os resultados da Tabela 5.2, o estimador dessa proporção sob AAS é dado por:

$$\hat{p}_d = n_d/n = \bar{d} \tag{7.5}$$

onde $n_d = \sum_{i \in s} d_i$ denota o tamanho da amostra no domínio d e $\bar{d}$ é a média amostral da variável derivada d_i.

A estimação da proporção no domínio só é necessária quando o *tamanho do domínio* na população (N_d) é desconhecido. Quando este tamanho for conhecido, não é necessário estimar a proporção p_d, e essa informação será útil também para melhorar estimativas de outros parâmetros do domínio, como veremos adiante.

A variância do estimador da proporção é dada por

$$V_{AAS}(\hat{p}_d) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N}{N-1} p_d(1-p_d) \tag{7.6}$$

e pode ser estimada sem vício usando

$$\widehat{V}_{AAS}(\hat{p}_d) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n}{n-1} \hat{p}_d(1-\hat{p}_d) \tag{7.7}$$

Ainda considerando o caso em que o *tamanho do domínio* na população é desconhecido, o próximo parâmetro do domínio que podemos estimar é o total Y_d. O estimador não viciado deste total é obtido simplesmente aplicando o estimador de Horvitz-Thompson à variável derivada y_{id}, resultando em:

$$\widehat{Y}_d = N/n \sum_{i \in s} y_{id} = N \frac{t_d}{n} \tag{7.8}$$

onde t_d é a soma amostral da variável y para unidades pertencentes ao domínio d, dada por $t_d = \sum_{i \in s} y_{id}$.

A variância deste estimador é obtida diretamente dos resultados usuais da estimação de totais sob AAS para a variável derivada y_{id}, levando a

$$V_{AAS}(\widehat{Y}_d) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{i \in U} \left(y_{id} - \frac{Y_d}{N}\right)^2 \tag{7.9}$$

e pode ser estimada de forma não viciada usando

$$\widehat{V}_{AAS}(\widehat{Y}_d) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{i \in s} \left(y_{id} - \frac{t_d}{n}\right)^2 \tag{7.10}$$

A média no domínio $\overline{Y}_d$ pode ser estimada usando:

$$\overline{y}_d = \frac{\widehat{Y}_d}{\widehat{N}_d} = \frac{(N/n)\sum_{i \in s} y_{id}}{(N/n)\sum_{i \in s} d_i} = \sum_{i \in s} y_{id}/n_d \tag{7.11}$$

Note que o estimador da média do domínio d é uma razão de totais estimados, e que n_d é uma *variável aleatória*, aparecendo no denominador da expressão do estimador da média no domínio.

Para obter sua variância, podemos recorrer aos resultados para estimação de razões disponíveis no capítulo 6, resultando em:

$$\begin{aligned} V_{AAS}(\overline{y}_d) &\cong \frac{1}{p_d^2}\left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{i \in U} \left(y_{id} - \overline{Y}_d d_i\right)^2 \\ &= \frac{1}{p_d^2}\left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{i \in U} d_i \left(y_i - \overline{Y}_d\right)^2 \end{aligned} \tag{7.12}$$

O estimador de variância correspondente é dado por:

$$\widehat{V}_{AAS}(\overline{y}_d) = \frac{1}{\hat{p}_d^2}\left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{i \in s} d_i \left(y_i - \overline{y}_d\right)^2 \tag{7.13}$$

7.4 Intervalos de confiança para os parâmetros para o domínio

A obtenção de intervalos de confiança na estimação de parâmetros populacionais para o domínio requer *amostra grande no domínio*, isto é, requer que n_d seja suficientemente grande para justificar o uso da aproximação pela distribuição Normal para os vários estimadores considerados. Quando este for o caso, valem as seguintes aproximações:

$$\left(\hat{p}_d - p_d\right) / \sqrt{\widehat{V}_{AAS}\left(\hat{p}_d\right)} \approx N(0;1) \text{ para } n_d \text{ grande.}$$

$$\left(\widehat{Y}_d - Y_d\right) / \sqrt{\widehat{V}_{AAS}\left(\widehat{Y}_d\right)} \approx N(0;1) \text{ para } n_d \text{ grande.}$$

$$\left(\overline{y}_d - \overline{Y}_d\right) / \sqrt{\widehat{V}_{AAS}(\overline{y}_d)} \approx N(0;1) \text{ para } n_d \text{ grande.}$$

Desse modo, as respectivas expressões dos intervalos de confiança de nível (1-α)% para a estimação de total e média do domínio d sob AAS passam a ser:

$$IC_{AAS}(p_d; 1-\alpha) = \left[\hat{p}_d \mp z_{\alpha/2}\widehat{V}_{AAS}(\hat{p}_d)\right] \tag{7.14}$$

$$IC_{AAS}(Y_d; 1-\alpha) = \left[\widehat{Y}_d \mp z_{\alpha/2}\sqrt{\widehat{V}_{AAS}\left(\widehat{Y}_d\right)}\right] \tag{7.15}$$

$$IC_{AAS}(\overline{Y}_d; 1-\alpha) = \left[\overline{y}_d \mp z_{\alpha/2}\sqrt{\widehat{V}_{AAS}\left(\overline{y}_d\right)}\right] \tag{7.16}$$

7.5 Propriedades condicionais (fixando n_d)

Cochran (1977) (seção 2.12) sugere analisar a distribuição da média considerando o tamanho da amostra no domínio n_d como fixado em seu valor observado. Nesse caso, mostra que as n_d observações na amostra s formam uma AAS da população U_d. Segue-se então que:

$$V_{AAS}(\overline{y}_d | \, n_d > 0) = \left(\frac{1}{n_d} - \frac{1}{N_d}\right) S_d^2 \tag{7.17}$$

A variância pode então ser estimada usando:

$$\widehat{V}_{AAS}(\overline{y}_d | \, n_d > 0) = \left(\frac{1}{n_d} - \frac{1}{N_d}\right) s_d^2 \tag{7.18}$$

onde $s_d^2 = \sum_{i \in s} d_i (y_i - \overline{y}_d)^2 / (n_d - 1)$ é um estimador não viciado para S_d^2 (quando n_d é fixado).

Quando N_d é desconhecido, $\widehat{V}_{AAS}(\overline{y}_d | \, n_d > 0)$ não é calculável.

Cochran (1977) (p. 35) sugere substituir a fração amostral no domínio n_d/N_d pela fração amostral esperada no domínio, dada por n/N, o que implica em:

$$\widehat{V}^*_{AAS}(\overline{y}_d | \, n_d > 0) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n_d} \tag{7.19}$$

7.6 Estimação do total no domínio quando tamanho do domínio é conhecido

Em algumas situações de pesquisa, é possível supor conhecido o *tamanho do domínio* N_d, a partir de dados de uma fonte secundária confiável. Nesses casos, temos as seguintes consequências:

a) Não é necessário estimar a proporção de unidades no domínio p_d;

b) Fica disponível um estimador alternativo para o total populacional Y_d que pode ser mais preciso que o estimador tipo Horvitz-Thompson descrito na expressão (7.8).

Para obter o estimador alternativo, note que:

$$Y_d = N_d \overline{Y}_d \tag{7.20}$$

Logo, é possível usar o *estimador tipo razão* para o total dado por:

$$\widehat{Y}_d^R = N_d \overline{y}_d = N_d / n_d \sum_{i \in s} y_{id} \tag{7.21}$$

A variância condicional deste estimador, considerando fixado o tamanho da amostra no domínio d, é dada por:

$$V_{AAS}\left(\widehat{Y}_d^R | \, n_d > 0\right) = N_d^2 \left(\frac{1}{n_d} - \frac{1}{N_d}\right) S_d^2 \tag{7.22}$$

Essa variância pode ser estimada usando:

$$\widehat{V}_{AAS}\left(\widehat{Y}_d^R | \, n_d > 0\right) = N_d^2 \left(\frac{1}{n_d} - \frac{1}{N_d}\right) s_d^2 \tag{7.23}$$

A Tabela 7.1 apresenta um resumo da estimação de parâmetros proporção, média e total do domínio d sob AAS.

Tabela 7.1: Estimadores dos parâmetros proporção, média e total do domínio d sob AAS

Parâmetro do Domínio d	**Estimador do Domínio d sob AAS**
$p_d = \sum_{i \in U} d_i / N = N_d / N$	$\hat{p}_d = n_d / n \ (N_d \text{ desconhecido})$
$\overline{Y}_d = Y_d / N_d$	$\overline{y}_d = \widehat{Y}_d / \widehat{N}_d = \sum_{i \in s} y_{id} / n_d \ (N_d \text{ desconhecido})$
$Y_d = N_d \overline{Y}_d$	$\widehat{Y}_d = N/n \sum_{i \in s} y_{id} \ (N_d \text{ desconhecido})$ $\widehat{Y}_d^R = N_d \overline{y}_d \ (N_d \text{ conhecido})$
$V_{AAS}(\hat{p}_d)$	$\widehat{V}_{AAS}(\hat{p}_d) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n}{n-1} \hat{p}_d (1 - \hat{p}_d)$
$V_{AAS}(\overline{y}_d)$	$\widehat{V}_{AAS}(\overline{y}_d) = \frac{1}{\hat{p}_d^2} \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{n-1} \sum_{i \in s} d_i \left(y_i - \overline{y}_d\right)^2$
$V_{AAS}(\overline{y}_d \mid n_d > 0, \text{ fixo})$	$\widehat{V}_{AAS}(\overline{y}_d \mid n_d > 0) = \left(1 - \frac{n_d}{N_d}\right) \frac{s_d^2}{n_d}$ $\widehat{V}^*_{AAS}(\overline{y}_d \mid n_d > 0) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n_d} \ (N_d \text{ desconhecido})$
$V_{AAS}(\widehat{Y}_d)$	$\widehat{V}_{AAS}(\widehat{Y}_d) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{N-1} \sum_{i \in s} \left(y_{id} - \frac{t_d}{n}\right)^2$
$V_{AAS}\left(\widehat{Y}_d^R \mid n_d > 0\right)$	$\widehat{V}_{AAS}\left(\widehat{Y}_d^R \mid n_d > 0\right) = N_d^2 \left(1 - \frac{n_d}{N_d}\right) \frac{s_d^2}{n_d}$

Exemplo 7.1 Estimando totais e médias por domínios

Considere os dados da população de municípios brasileiros fornecidos no arquivo ‘MunicBR_dat.rds’.

1. Selecione uma AAS de $n = 250$ municípios, e use esta amostra para estimar os seguintes parâmetros populacionais:

    a. População total por macro-região do Brasil, e correspondentes margens de erro relativo ao nível de confiança de 95%; suponha *conhecidos* os tamanhos dos domínios;

b. População total por macro-região do Brasil, e correspondentes margens de erro relativo ao nível de confiança de 95%; suponha *desconhecidos* os tamanhos dos domínios;

c. Média da densidade habitacional por km^2 por município para municípios com população igual ou superior a 100 mil habitantes, e correspondente margem de erro ao nível de confiança de 95%. Nesse caso, considere que o tamanho do domínio relevante é *desconhecido*.

2. Usando os dados populacionais, calcule os valores dos parâmetros estimados em a) e c) e compare com suas estimativas amostrais. Comente.

Solução do Exemplo 7.1 usando R

```
# Limpa área de trabalho
rm(list = ls())

# Carrega biblioteca(s) requerida(s)
library(sampling)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Leitura dos dados
MunicBR_dat <- readRDS(file="MunicBR_dat.rds")
str(MunicBR_dat)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    5570 obs. of  6 variables:
##  $ CodMunic : chr  "1100015" "1100023" "1100031" "1100049" ...
##  $ SiglaUF  : chr  "RO" "RO" "RO" "RO" ...
##  $ CodUF    : chr  "11" "11" "11" "11" ...
##  $ Pop      : num  25728 101269 6495 85863 18041 ...
##  $ Area     : num  7067 4427 1314 3793 2783 ...
##  $ Densidade: num  3.64 22.88 4.94 22.64 6.48 ...
```

```
# Cria código de região nos dados da população
reglabels = c("Norte", "Nordeste", "Sudeste", "Sul", "Centro-Oeste")
MunicBR_dat <- mutate(MunicBR_dat,
                      Regiao = substr(CodUF,1,1))  %>%
               mutate(Regiao = factor(Regiao, labels=reglabels))
table(MunicBR_dat$Regiao)
```

```
##
##        Norte     Nordeste      Sudeste          Sul Centro-Oeste
##          450         1794         1668         1191          467
```

```
# Define senha para geração de números aleatórios para permitir repetição
set.seed(123)

# Item 1
# Tamanho da amostra
(n <- 250)
```

```
## [1] 250
```

```
# Tamanho da população
(N <- nrow(MunicBR_dat))
```

```
## [1] 5570
```

```r
# Seleciona amostra AAS dos municípios
munic_amo <- getdata(MunicBR_dat, srswor(n,N))
str(munic_amo)
```

```
## 'data.frame':    250 obs. of  8 variables:
##  $ ID_unit  : int  4 57 137 234 253 254 256 283 313 332 ...
##  $ CodMunic : chr  "1100049" "1200179" "1400027" "1505205" ...
##  $ SiglaUF  : chr  "RO" "AC" "RR" "PA" ...
##  $ CodUF    : chr  "11" "12" "14" "15" ...
##  $ Pop      : num  85863 9836 10432 30088 17774 ...
##  $ Area     : num  3793 1703 28472 3852 4115 ...
##  $ Densidade: num  22.638 5.777 0.366 7.81 4.32 ...
##  $ Regiao   : Factor w/ 5 levels "Norte","Nordeste",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
table(munic_amo$Regiao)
```

```
##
##        Norte     Nordeste      Sudeste          Sul Centro-Oeste
##           11           87           81           46           25
```

```r
# Soluções para item 1
# Estima população total por macro-região
# Calcula as contagens populacionais por domínio
(N_d <- MunicBR_dat %>%
        group_by(Regiao) %>%
        summarise(N_d = n()) %>%
        select(Regiao, N_d) )
```

```
## # A tibble: 5 x 2
##   Regiao         N_d
##   <fct>        <int>
## 1 Norte          450
## 2 Nordeste      1794
## 3 Sudeste       1668
## 4 Sul           1191
## 5 Centro-Oeste   467
```

```
# Calcula médias amostrais por domínio
(ybar_d <- munic_amo %>%
    group_by(Regiao) %>%
    summarise(ybar_d = mean(Pop),
              t_d = sum(Pop)) %>%
    select(Regiao, ybar_d, t_d) )
```

```
## # A tibble: 5 x 3
##   Regiao       ybar_d     t_d
##   <fct>         <dbl>   <dbl>
## 1 Norte        24506.  269567
## 2 Nordeste     32776. 2851482
## 3 Sudeste      53116. 4302358
## 4 Sul          23934. 1100956
## 5 Centro-Oeste  9297.  232419
```

```
# Junta informações das duas tabelas por Região
resumo_regiao <- left_join(N_d, ybar_d, by=c("Regiao"))
# a. Calcula estimativas dos totais, supondo contagens populacionais conhecidas
(Total_reg_est1 <- resumo_regiao %>%
   mutate(Total_reg_est1 = N_d * ybar_d) %>%
   select(Regiao, Total_reg_est1))
```

```
## # A tibble: 5 x 2
##   Regiao       Total_reg_est1
##   <fct>                 <dbl>
## 1 Norte             11027741.
## 2 Nordeste          58799525.
## 3 Sudeste           88596705.
## 4 Sul               28505187.
## 5 Centro-Oeste       4341587.
```

```
# b. Calcula estimativas dos totais, supondo contagens populacionais desconhecidas
(Total_reg_est2 <- resumo_regiao %>%
    mutate(Total_reg_est2 = N * t_d / n) %>%
    select(Regiao, Total_reg_est2))
```

```
## # A tibble: 5 x 2
##   Regiao       Total_reg_est2
##   <fct>                 <dbl>
## 1 Norte              6005953.
## 2 Nordeste          63531019.
## 3 Sudeste           95856536.
## 4 Sul               24529300.
## 5 Centro-Oeste       5178295.
```

```
# Calcula totais populacionais para comparação
(Total_reg_pop <- MunicBR_dat %>%
    group_by(Regiao) %>%
    summarise(Total_reg_pop = sum(Pop)) %>%
    select(Regiao, Total_reg_pop))
```

```
## # A tibble: 5 x 2
##   Regiao       Total_reg_pop
##   <fct>                <dbl>
## 1 Norte             17013559
## 2 Nordeste          55794707
## 3 Sudeste           84465570
## 4 Sul               28795762
## 5 Centro-Oeste      14993191
```

```
# Estima densidade habitacional média por km2 no Brasil
(r_chapeu <- munic_amo %>%
             summarise(Popm = mean(Pop),
                       Aream = mean(Area)) %>%
             mutate(Densm = Popm / Aream) %>%
             select(Densm))
```

```
##      Densm
## 1 34.16426
```

```
# c. Estima média da densidade habitacional por km2 por município
(media.densidade <- munic_amo %>%
                    summarise(Densm = mean(Densidade)))
```

```
##      Densm
## 1 135.0281
```

```r
# Adiciona valor de r_chapeu aos dados da amostra
munic_amo <- cbind(munic_amo, r_chapeu)

# Calcula precisão
(precisao.r_chapeu <- munic_amo %>%
                      mutate(Z = Pop - Densm * Area) %>%
                      summarise(varZ = var(Z),
                                Aream = mean(Area),
                                Densm = mean(Densm)) %>%
                      mutate(dp.r_chapeu = sqrt((1/n - 1/N)*varZ)/Aream,
                             cv.r_chapeu = 100 * dp.r_chapeu / Densm)  %>%
                      select(dp.r_chapeu, cv.r_chapeu))
```

```
##   dp.r_chapeu cv.r_chapeu
## 1    7.630939    22.33603
```

```r
# Calcula precisão do item c)
(precisao.media.densidade <- munic_amo %>%
         summarise(Densv = var(Densidade),
                   Densm = mean(Densidade)) %>%
         mutate(dp.media.dens = sqrt((1/n - 1/N)*Densv),
                cv.media.dens = 100 * dp.media.dens / Densm) %>%
         select(dp.media.dens, cv.media.dens))
```

```
##   dp.media.dens cv.media.dens
## 1      38.14405      28.24896
```

```r
# Calcula densidade habitacional média por km2 no Brasil
(R <- MunicBR_dat %>%
      summarise(Popm = mean(Pop),
                Aream = mean(Area)) %>%
      mutate(Densidade_pop = Popm / Aream) %>%
      select(Densidade_pop))
```

```
## # A tibble: 1 x 1
##   Densidade_pop
##           <dbl>
## 1          23.6
```

```
# Estima média da densidade habitacional por km2 por município
(densidade_media_pop <- MunicBR_dat %>%
    summarise(densidade_media_pop = mean(Densidade)))
```

```
## # A tibble: 1 x 1
##   densidade_media_pop
##                 <dbl>
## 1                114.
```

7.7 Comparação da eficiência dos estimadores de total do domínio

Foram propostos dois estimadores para o total populacional Y_d, sendo um deles viável apenas quando o *tamanho do domínio* é conhecido. Nesse caso, é importante saber se o estimador tipo razão é preferível ao estimador tipo Horvitz-Thompson em termos de eficiência. Comparando as respectivas variâncias, de acordo com Cochran (1977) (p. 38), tem-se que a eficiência relativa do estimador tipo razão é dada por:

$$\frac{V_{AAS}\left(\widehat{Y}_d^R \,|\, n_d > 0\right)}{V_{AAS}\left(\widehat{Y}_d \,|\, n_d > 0\right)} \cong \frac{S_d^2}{S_d^2 + (1-p_d)\overline{Y}_d^2} = \frac{C_d^2}{C_d^2 + (1-p_d)} \tag{7.24}$$

onde $C_d^2 = S_d^2/\overline{Y}_d^2$ é a variância relativa da característica y no domínio d.

Isto mostra que conhecer o valor de N_d sempre melhora a eficiência do estimador de total do domínio d, e que a melhoria é maior quando p_d é pequena. Essa expressão também torna evidente que, para domínios raros na população, isto é, aqueles com valores bem pequenos de p_d, o estimador do total do domínio d para N_d desconhecido é bem pouco preciso, devendo talvez indicar a necessidade de uso de plano amostral que favoreça a estimação de Y_d com melhor precisão que uma AAS.

7.8 Estimação de proporções dentro do domínio

Em algumas situações práticas, o parâmetro de interesse é a proporção de unidades do domínio d que possuem um atributo ou característica A. Por exemplo, quando se deseja estimar a proporção de mulheres de 10 anos ou mais que já tiveram pelo menos um filho, ou quando se procura estimar a proporção de homens de 18 anos ou mais que prestaram o serviço militar. Em casos como os acima citados, o problema é estimar proporções dos domínios da população: mulheres de 10 anos ou mais; e homens com 18 anos ou mais.

Nesses casos, a variável de pesquisa y seria dada por:

$$y_i = I(i \in A) = \begin{cases} 1 & \text{se } \ i \text{ possui o atributo A,} \\ 0 & \text{caso contrário.} \end{cases}$$

Na população como um todo, a proporção de unidades com atributo A é definida como $p = N_A/N$, e a estimação desta proporção foi tratada no capítulo 5. Considere a notação a seguir.

O número de unidades no domínio d que também possuem o atributo A é definido como:

$$N_{dA} = \sum_{i \in U} y_{id} = \sum_{i \in U} y_i d_i = \sum_{i \in U_d} y_i$$

E a proporção de unidades no domínio d que também possuem o atributo A é definida como:

$$p_{dA} = N_{dA}/N_d \tag{7.25}$$

Sob AAS, o estimador simples para p_{dA} pode ser obtido a partir do estimador(7.11), lembrando que agora a variável y é também do tipo indicadora, resultando em:

$$\hat{p}_{dA} = \sum_{i \in s} y_{id}/n_d = \frac{n_{dA}}{n_d} \tag{7.26}$$

onde n_{dA} denota o número de unidades na amostra do domínio d que também possuem o atributo A.

Considerando fixado o tamanho da amostra no domínio d, a variância condicional do estimador $\hat{p}_{dA}$ é dada por:

$$V_{AAS}(\hat{p}_{dA}) = \left(\frac{1}{n_d} - \frac{1}{N_d}\right) \frac{N_d}{N_d - 1} p_{dA}(1 - p_{dA}) \tag{7.27}$$

Um estimador da variância de $\hat{p}_{dA}$ sob AAS resulta em:

$$\widehat{V}_{AAS}(\hat{p}_{dA}) = \left(\frac{1}{n_d} - \frac{1}{N_d}\right)\hat{p}_{dA}(1-\hat{p}_{dA}) \tag{7.28}$$

Caso N_d não seja conhecido, a fração amostral no domínio n_d/N_d pode ser aproximado por n/N na expressão anterior, levando ao estimador

$$\widehat{V}_{AAS}(\hat{p}_{dA}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}_{dA}(1-\hat{p}_{dA})}{n_d} \tag{7.29}$$

Para completar a inferência sobre uma proporção de unidades portadoras do atributo A no domínio d, admite-se a validade da aproximação normal para a distribuição de $\hat{p}_{dA}$, tal como no capítulo 5, e se agrega uma *correção de continuidade*. Assim a expressão do intervalo de confiança para a proporção populacional p_{dA} é dada por:

$$IC_{AAS}(p_{dA}\,;\,1-\alpha) = \left[\hat{p}_{dA} \mp \left(z_{\alpha/2}\sqrt{\widehat{V}_{AAS}(\hat{p}_{dA})} + \frac{1}{2n_d}\right)\right] \tag{7.30}$$

7.9 Exercícios

Exercício 7.1 A população total de uma cidade é de $N = 50.000$ pessoas. Uma amostra de fração amostral igual a 20% é selecionada aleatoriamente sem reposição. Na amostra 4.000 pessoas estão na força de trabalho, das quais 200 estão sem emprego.

a. Que proporção da força de trabalho está desempregada?
b. Qual o intervalo com 90% de confiança para essa proporção populacional?

Exercício 7.2 Um economista deseja estimar o gasto médio com alimentação das famílias com crianças em uma determinada localidade conhecida como uma área onde residem famílias pobres. Está disponível uma listagem com os endereçoes da 250 famílias que habitam a localidade, porém é impossível identificar nesse cadastro as famílias com crianças. Foi selecionada e pesquisada uma AAS de 50 famílias. Destas, 42 famílias têm pelo menos uma criança. Foi apurado que o total do gasto semanal com alimentos dessas 42 famílias foi de $\$1720$ e que a soma dos quadrados dos gastos dessas mesmas famílias foi de 72000.

a. Estime a média semanal dos gastos com alimentação das famílias com criança da localidade.
b. Construa um intervalo de 95% de confiança para a média estimada.
c. Estime total semanal dos gastos com alimentação das famílias com criança da localidade.
d. Construa um intervalo de 95% de confiança para o total estimado.

Exercício 7.3 Considere a população de $N = 338$ fazendas produtoras de cana de açúcar fornecida no arquivo 'fazendas_dat.rds'.

Considere um plano AAS e tamanhos amostrais n variando no conjunto {5, 10, 20, 50, 100}. Imagine que há interesse em estimar dois parâmetros:

I. A média da variável Produtividade = Quant / Area;
II. A produtividade média por unidade de área na população.

Para cada um dos tamanhos de amostra considerados, realize as tarefas abaixo indicadas.

a. Obtenha 500 amostras por AAS da população de fazendas.
b. Use cada uma destas amostras para calcular estimativas dos dois parâmetros de interesse, e dos respectivos erros padrão.
c. Use as 500 estimativas pontuais obtidas para cada parâmetro para avaliar:
    - O comportamento do viés dos estimadores;
    - O comportamento do viés dos estimadores dos erros padrões;
    - A adequação da aproximação normal para a distribuição dos estimadores usados.

Exercício 7.4 Considere a população de $N = 338$ fazendas produtoras de cana de açúcar fornecida no arquivo 'fazendas_dat.rds'.

Considere agora um plano AAS e dois tamanhos amostrais $n_1 = 20$ e $n_2 = 100$. Agora o interesse é estimar a despesa com a *produção total de cana para um domínio de interesse*, definido como o conjunto de fazendas pertencentes às classes de tamanho 4, 5 e 6.

Para cada um dos tamanhos de amostra considerados, realize as tarefas abaixo indicadas.

a. Obtenha 500 amostras por AAS da população de fazendas.
b. Use cada uma destas amostras para calcular estimativas do parâmetro de interesse e dos correspondentes erros padrões supondo que o tamanho do domínio é *desconhecido*.
c. Use cada uma destas amostras para calcular estimativas do parâmetro de interesse e dos correspondentes erros padrões supondo que o tamanho do domínio é *conhecido*.
d. Use as 500 estimativas pontuais obtidas em b) e em c) para avaliar a adequação da aproximação normal para a distribuição dos estimadores usados, e também para avaliar qual dos dois estimadores fornece resultados mais precisos.

Exercício 7.5 Foi realizado um estudo sobre a distância percorrida pelos reponsáveis pelos domicílios de uma localidade de suas residencias até o seus locais de trabalho. Uma AAS de 30 responsáveis foi selecionada entre os 393 responsáveis por domicílios da área em questão. Durante as entrevistas foi verificado que alguns dos selecionados não precisavam

se deslocar para o trabalho por estarem aposentados, entre outras razões. Da amostra selecionada, apenas 24 dos responsáveis se adequavam ao estudo. A Tabela 7.2 da a distância percorrida por cada entrevistado para chegar ao trabalho.

Tabela 7.2: Distância percorrida para chegar ao trabalho pelos 24 entrevistados

8,5	10,2	25,1	5,0	6,3	7,9	15,8	2,1
9,2	4,2	8,3	4,2	6,7	10,1	15,6	22,1
10,0	6,1	7,9	1,5	8,0	11,0	20,2	9,3

a. Estime a média da distância entre as residências e os locais de trabalho dos responsáveis pelos domicílios da localidade.
b. Construa um intervalo de 95% de confiança para a média estimada.
c. Estime a distância total percorrida pelos responsáveis para chegar aos seus locais de trabalho.
d. Construa um intervalo de 95% de confiança para o total estimado.

Exercício 7.6 Uma cidade tem 468 escolas entre as quais uma amostra aleatória simples de 100 escolas foi selecionada. Verificou-se, após a seleção, que 54 eram escolas públicas e as demais, privadas. Foram pesquisados em cada escola o número de professores e o número de estudantes. Um resumo das informações é apresentado na Tabela 7.3.

Tabela 7.3: Resumo das informações amostrais para as 100 escolas da amostra

Tipo	n	$\sum y_i$	$\sum x_i$	$\sum y_i^2$	$\sum x_i^2$	$\sum y_i x_i$
Pública	54	31281	2024	29881219	1110901	1729349
Privada	46	13707	1075	6366785	33119	431041

a. Estime o total de alunos para as escolas públicas da cidade e construa um intervalo de 95% de confiança para a estimativa.
b. Estime o total de alunos para as escolas particulares da cidade e construa um intervalo de 95% de confiança para a estimativa.
c. Refaça o item a sabendo que existem 251 escolas públicas na cidade. Compare e comente os dois resultados.
d. Estime a razão (número de profesores)/(número de estudantes) para cada tipo de escola e calcule o respectivo erro padrão das estimativas.

Exercício 7.7 Um dentista, o Dr. A, selecionou uma amostra do tipo AAS de tamanho 20 entre 200 crianças de uma comunidade. Ele contou o número de dentes com cárie de cada uma das crianças e o resultado é mostrado na Tabela de frequências 7.4.

Tabela 7.4: Tabela de frequências do número de cáries por criança

Número de dentes com cárie	0	1	2	3	4	5	6	7	8	9	10
Número de crianças	8	4	2	2	1	1	0	0	0	1	1

Outro dentista, o Dr. B, examinou todas as 200 crianças e apenas verificou que 60 delas tinham, pelo menos, um dente com cárie.

a. Estime a média de dentes cariados na população de crianças utilizando apenas as informações obtidas pelo Dr. A, obtendo, também, um intervalo de 95% de confiança para o parâmetro estimado.
b. Repita o item a utilizando as informações obtidas pelos dois dentistas.
c. Estime a média de dentes com cárie somente para as crianças da população que tem dentes cariados, usando somente as informações do Dr. A. Dê um intervalo de 95% de confiança para a média.
d. Repita o item c utilizando as informações obtidas pelos dois dentistas.

Exercício 7.8 Utilizando o arquivo dos municípios brasileiros ($MunicBR\_dat.rds$), selecione uma AAS de 300 municípios. Utilizando as informações contidas nessa amostra:

a. Estime a área total de cada uma das 5 grandes regiões brasileiras (N, NE, SE, S e CO) e calcule a estimativa do CV para cada caso.
b. Estime a proporção de municípios brasileiros com população menor que 10000 habitantes e respectivo erro padrão.
c. Estime a população total, para o Brasil, dos municípios com menos de 10000 habitantes e de um intervalo de 95% de confiança.

Exercício 7.9 Selecione uma amostra de 30 fazendas do arquivo $fazendas\_dat.rds$, estime os seguintes parâmetros populacionais e respectivos erros padrão:

a. Razão entre Despesa e Receita para os estabelecimentos com área menor que 100.
b. Razão entre Despesa e Receita para os estabelecimentos com área igual ou maior que 100.
c. As razões podem ser consideradas iguais ao nível de significância $\alpha = 5\%$?

Exercício 7.10 Uma granja tem um plantel de 10000 aves de 5 espécies diferentes. Foi selecionada uma AAS de 100 aves. A Tabela 7.5 apresenta as contagens dos animais da

amostra.

Tabela 7.5: Quantidade de aves na amostra por espécie

Espécie	Quantidade
Pato	20
Galinha	21
Peru	22
Ganso	18
Marreco	19
Total	**100**

a. Estime o total de cada espécie de ave do plantel.

b. Construa um intervalo de 95% de confiança para cada espécie.