

# MAT02025 - Amostragem 1

**AAS: estimação de proporções para classificações em mais de duas categorias e proporções em subpopulações**

Rodrigo Citton P. dos Reis  
citton.padilha@ufrgs.br

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL  
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA  
DEPARTAMENTO DE ESTATÍSTICA

Porto Alegre, 2023

## Proporções: duas ou mais categorias

# Proporções: duas ou mais categorias

- ▶ Até o momento foi tratado o caso em que temos uma **variável dicotômica (ou dicotomizada)**, resultando na **classificação** da população em **duas categorias**.
- ▶ Muitas vezes temos a necessidade de definir uma **classificação** com **mais de duas categorias**.

## Proporções: duas ou mais categorias

Por exemplo,

- ▶ Estudar a distribuição por **faixas etárias** de um grupo de pessoas.
  - ▶ 18-24; 25-44; 45-64; 65+.
- ▶ Estudar a **classificação econômica das empresas** de determinado país.
  - ▶ Do Setor **primário; secundário; ou terciário**.
  - ▶ Empresa do tipo **extrativista; agropecuária; industrial; comercial; serviços**.
- ▶ Estimar a **intenção de voto** das chapas do DCE em uma **eleição com mais de 2 chapas**, além das possibilidades de voto em branco ou nulo ou, ainda, eleitores indecisos.
  - ▶ Intenção de voto na **Chapa 1; Chapa 2; ... ; Chapa 6**.

Nesses casos, há interesse de estimar a **proporção de unidades em cada uma das possíveis categorias** e a **respectiva precisão**.

## Proporções: duas ou mais categorias

**Exemplo:** seja uma escola com 1.000 alunos distribuídos entre as 9 etapas do ensino fundamental.

Etapa de ensino	Alunos	Proporção
1º ano	110	0,110
2º ano	108	0,108
3º ano	110	0,110
4º ano	115	0,115
5º ano	104	0,104
6º ano	119	0,119
7º ano	116	0,116
8º ano	107	0,107
9º ano	111	0,111
<b>Total</b>	<b>1.000</b>	<b>1,000</b>

## Proporções: duas ou mais categorias

Para este exemplo, podemos pensar na seguinte notação:

- ▶  $X$  é uma característica (um atributo, ou classificação) que assume os seguintes valores

$$X = \begin{cases} C_1, & \text{se aluno do 1º ano} \\ C_2, & \text{se aluno do 2º ano} \\ \vdots & \vdots \\ C_9, & \text{se aluno do 9º ano} \end{cases}$$

- ▶ Gostaríamos de estimar  $P_1, P_2, \dots, P_9$ , em que  $P_j$  é a proporção populacional de unidades na classe  $C_j$ .

## Proporções: duas ou mais categorias

- ▶ Observe que, para calcular as proporções em cada uma das categorias, na verdade o que se faz é atribuir o valor **1** às unidades da categoria em questão e o valor **0** para as unidades pertencentes às demais categorias.
- ▶ Em outras palavras, se a variável tem  **$m$**  categorias é como se fossem  **$m$**  problemas com duas categorias.
- ▶ A proporção de unidades da população pertencentes à categoria  $C \in \{1, 2, \dots, m\}$ , é dada por:

$$P_C = \frac{A_C}{N},$$

em que  $A_C$  é o **número de unidades na categoria  $C$  na população** e  $N$  é o tamanho total da população.

## Proporções: duas ou mais categorias

- ▶ Seja uma amostra aleatória simples de tamanho  $n$  e seja a variável indicadora  $Y_i$  definida como:

$$Y_i = \begin{cases} 1, & \text{se a unidade } i \text{ pertence à categoria } C; \\ 0, & \text{se a unidade } i \text{ pertence a outra categoria.} \end{cases}$$

- ▶ Com tal definição pode-se ver que o **número de unidades da categoria  $C$  na amostra** será dado por:

$$a_C = \sum_{i=1}^n Y_i, \quad C \in \{1, 2, \dots, m\}.$$



## Proporções: duas ou mais categorias

- Um estimador para a proporção de unidades populacionais pertencentes à categoria  $C$  é dado por:

$$p_C = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{a_C}{n}, \quad C \in \{1, 2, \dots, m\}.$$

# Proporções: duas ou mais categorias

## Observações

- ▶ O problema foi reduzido ao caso de estimar proporções em variáveis com duas categorias.
- ▶ Pode-se obter, também, estimativas de precisão utilizando os mesmos resultados já apresentados nas aulas 17, 18 e 19.

## Agrupando categorias

## Agrupando categorias

- ▶ Muitas vezes pode-se estar interessado em estimar **proporções para agrupamentos das categorias originais**.
- ▶ Voltando ao **exemplo da escola do ensino fundamental**, pode ser de interesse estudar a **proporção** de seus **alunos** que estão **matriculados no primeiro segmento do ensino fundamental** (1º até o 5º ano).
  - ▶ Nesse caso, seriam contabilizados como pertencentes à categoria **C** de interesse todos os alunos do 1º até o 5º ano, para os quais  **$Y = 1$** , sendo  **$Y = 0$**  para os demais alunos da escola.

## Não-resposta

- ▶ Outro caso de interesse ocorre quando, na aplicação de um questionário, por exemplo, aparecem **respondentes que se recusaram a responder** ou, mesmo, disseram que **não sabiam a resposta**.
- ▶ Num caso como esse, pode-se estar interessado em estimar a **proporção das pessoas que responderam determinada alternativa**, entre as pessoas que efetivamente responderam a pesquisa escolhendo uma das alternativas válidas (respostas válidas, votos válidos, etc.).

# Não-resposta

- ▶ Um exemplo prático seria uma pesquisa sobre a intenção de voto numa eleição com apenas duas chapas.
  - ▶ Nesse caso, o entrevistado poderia responder que votará na chapa A, na chapa B, que votará **nulo** ou *em branco*, ou **não sabe** em que chapa irá votar, onde apenas as duas primeiras alternativas seriam consideradas como **votos válidos**.

## Não-resposta

- Pode-se estimar a proporção para cada uma das cinco categorias iniciais (**A**, **B**, **nulo**, *em branco*, ou **não sabe**) ou apenas a **proporção de votos válidos** para cada uma das duas chapas:

$$p_A = \frac{a_A}{a_A + a_B} \quad \text{e} \quad p_B = \frac{a_B}{a_A + a_B}.$$

- Vale notar que na expressão acima, tanto o numerador como o denominador do estimador da proporção são variáveis aleatórias, pois a população (eleitores que efetivamente vão votar em uma das duas chapas) é desconhecida.

## Um exemplo



# Índice de Desempenho Acadêmico

- ▶ O **Índice de Desempenho Acadêmico da Califórnia** (*Academic Performance Index*, API) é calculado a partir de testes padronizados administrados a alunos em escolas da Califórnia.
- ▶ Além dos dados de desempenho acadêmico das escolas, há uma ampla gama de variáveis socioeconômicas disponíveis.
- ▶ Os dados a seguir se referem a uma amostra aleatória simples de tamanho  $n = 200$  de uma população de  $N = 6194$  escolas.

# Índice de Desempenho Acadêmico

```
library(survey)
data(api)

# ?apisrs
head(apisrs[,1:4])
```

##	cds	stype	name	sname
## 1039	15739081534155	H	McFarland High	McFarland High
## 1124	19642126066716	E	Stowers (Cecil	Stowers (Cecil B.) Elementary
## 2868	30664493030640	H	Brea-Olinda Hig	Brea-Olinda High
## 1273	19644516012744	E	Alameda Element	Alameda Elementary
## 4926	40688096043293	E	Sunnyside Eleme	Sunnyside Elementary
## 2463	19734456014278	E	Los Molinos Ele	Los Molinos Elementary

# Índice de Desempenho Acadêmico

- ▶ As escolas podem ser classificadas como “**Elementary**” (E), “**Middle**” (M), “**High School**” (H).

```
(a <- table(apisrs$type)) # a_E, a_H, a_M
```

```
##  
##      E      H      M  
## 142    25    33
```

```
(p <- prop.table(a)) # p_E, p_H, p_M
```

```
##  
##      E      H      M  
## 0.710 0.125 0.165
```

# Índice de Desempenho Acadêmico

- Podes obter estimativas do erro padrão das proporções.

```
n <- 200
N <- 6194
f <- n/N
(q <- 1 - p) # q_E, q_H, q_M
```

```
##
##      E      H      M
## 0.290 0.875 0.835
```

```
round(ep <- sqrt( (1 - f) * ((p*q)/(n - 1)) ), 4)
```

```
##
##      E      H      M
## 0.0316 0.0231 0.0259
```

# Índice de Desempenho Acadêmico

- E também podemos obter limites de confiança (**IC 95% usando a aproximação normal**).

```
round(li <- p - qnorm(p = 0.975) * ep , 3)
```

```
##  
##      E      H      M  
## 0.648 0.080 0.114
```

```
round(ls <- p + qnorm(p = 0.975) * ep + 1/(2*n) , 3)
```

```
##  
##      E      H      M  
## 0.775 0.173 0.218
```

# Índice de Desempenho Acadêmico

- Uma outra forma (mais direta e confiável) de se estimar estas proporções populacionais com os respectivos intervalos de confiança é utilizando o pacote survey.

```
# O objeto design
api.des <- svydesign(id = ~1,
                   fpc = ~fpc,
                   data = apisrs)
```

# Índice de Desempenho Acadêmico

```
# Estimativa da proporção de escolas  
# do tipo "Elementary/Middle/High School"  
# e intervalo de confiança de 95%  
svyciprop(formula = ~I(stype == "E"),  
           design = api.des, method = "mean")
```

```
##                2.5% 97.5%  
## I(stype == "E") 0.710 0.648 0.77
```

```
svyciprop(formula = ~I(stype == "H"),  
           design = api.des, method = "mean")
```

```
##                2.5% 97.5%  
## I(stype == "H") 0.1250 0.0795 0.17
```

```
svyciprop(formula = ~I(stype == "M"),  
           design = api.des, method = "mean")
```

```
##                2.5% 97.5%  
## I(stype == "M") 0.165 0.114 0.22
```

# Índice de Desempenho Acadêmico

- ▶ A função `svyciprop` também possui outros métodos de construção dos ICs de  $P$ .

*# Método xlogit*

```
svyciprop(formula = ~I(stype == "E"),
          design = api.des, method = "xlogit")
```

```
##                2.5% 97.5%
## I(stype == "E") 0.710 0.644  0.77
```

```
svyciprop(formula = ~I(stype == "H"),
          design = api.des, method = "xlogit")
```

```
##                2.5% 97.5%
## I(stype == "H") 0.1250 0.0861  0.18
```

```
svyciprop(formula = ~I(stype == "M"),
          design = api.des, method = "xlogit")
```

```
##                2.5% 97.5%
## I(stype == "M") 0.165 0.120  0.22
```



# Índice de Desempenho Acadêmico

```
# Método likelihood
```

```
svyciprop(formula = ~I(stype == "E"),  
           design = api.des, method = "likelihood")
```

```
##                               2.5% 97.5%  
## I(stype == "E") 0.710 0.645  0.77
```

```
svyciprop(formula = ~I(stype == "H"),  
           design = api.des, method = "likelihood")
```

```
##                               2.5% 97.5%  
## I(stype == "H") 0.1250 0.0842  0.18
```

```
svyciprop(formula = ~I(stype == "M"),  
           design = api.des, method = "likelihood")
```

```
##                               2.5% 97.5%  
## I(stype == "M") 0.165 0.118  0.22
```

# Índice de Desempenho Acadêmico

- ▶ O coeficiente de confiança dos ICs também podem ser especificados pela função `svyciprop`.

```
# IC 90%
```

```
svyciprop(formula = ~I(stype == "E"), design = api.des,  
           method = "mean", level = 0.9)
```

```
##                               5%  95%
```

```
## I(stype == "E") 0.710 0.658 0.76
```

```
svyciprop(formula = ~I(stype == "H"), design = api.des,  
           method = "xlogit", level = 0.9)
```

```
##                               5%  95%
```

```
## I(stype == "H") 0.1250 0.0916 0.17
```

```
svyciprop(formula = ~I(stype == "M"), design = api.des,  
           method = "xlogit", level = 0.9)
```

```
##                               5%  95%
```

```
## I(stype == "M") 0.165 0.127 0.21
```

# Índice de Desempenho Acadêmico

```
# IC 99%
```

```
svyciprop(formula = ~I(stype == "E"), design = api.des,  
           method = "mean", level = 0.99)
```

```
##                                0.5% 99.5%  
## I(stype == "E") 0.710 0.628  0.79
```

```
svyciprop(formula = ~I(stype == "H"), design = api.des,  
           method = "xlogit", level = 0.99)
```

```
##                                0.5% 99.5%  
## I(stype == "H") 0.1250 0.0763  0.2
```

```
svyciprop(formula = ~I(stype == "M"), design = api.des,  
           method = "xlogit", level = 0.99)
```

```
##                                0.5% 99.5%  
## I(stype == "M") 0.165 0.108  0.24
```

# Estimação de proporções dentro de setores

## Estimação de proporções dentro de setores

- ▶ Em algumas situações práticas, o parâmetro de interesse é a proporção de unidades no **setor**<sup>1</sup>  $j$  que possuem um atributo ou característica  $C$ .
  - ▶ Por exemplo, quando se deseja estimar a **proporção** de **mulheres de 15 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 nos setores da população:
  - ▶ **mulheres de 15 anos ou mais**;
  - ▶ e **homens de 18 anos ou mais**.
- ▶ **Pergunta:** qual o atributo associado ao parâmetro de proporção que queremos estimar?

---

<sup>1</sup>domínio, subgrupo ou subpopulação.

# Estimação de proporções dentro de setores

Rev Saude Publica. 2017;51 Supl 1:125

Suplemento DCNT e Inquéritos  
Artigo Original



<http://www.rsp.fsp.usp.br/>

Revista de  
Saúde Pública

## Fatores associados ao diabetes autorreferido segundo a Pesquisa Nacional de Saúde, 2013

Deborah Carvalho Malta<sup>I</sup>, Regina Tomie Ivata Bernal<sup>II</sup>, Betine Pinto Moehleck Iser<sup>III,IV</sup>,  
Célia Landmann Szwarcwald<sup>V</sup>, Bruce Bartholow Duncan<sup>III</sup>, Maria Inês Schmidt<sup>IV</sup>

<sup>I</sup> Departamento de Enfermagem Materno Infantil e Saúde Pública. Escola de Enfermagem. Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brasil

<sup>II</sup> Núcleo de Pesquisas Epidemiológicas em Nutrição e Saúde. Universidade de São Paulo. São Paulo, SP, Brasil

<sup>III</sup> Programa de Pós-Graduação em Epidemiologia. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS, Brasil

<sup>IV</sup> Faculdade de Medicina. Universidade do Sul de Santa Catarina. Tubarão, SC, Brasil

<sup>V</sup> Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. Fundação Oswaldo Cruz. Rio de Janeiro, RJ, Brasil

# Estimação de proporções dentro de setores

**Tabela 1.** Prevalência de diabetes em adultos por sexo, segundo fatores sociodemográficos. Pesquisa Nacional de Saúde, Brasil, 2013.

Variável	Total		Masculino		Feminino	
	%	IC95%	%	IC95%	%	IC95%
Total	6,2	5,9–6,6	5,4	4,8–5,9	7	6,5–7,5
Idade (anos)						
18–24	0,5	0,3–0,8	0,4	0,1–0,7	0,6	0,2–1,1
25–34	0,8	0,6–1,1	0,8	0,4–1,2	0,9	0,6–1,2
35–44	3	2,4–3,5	2,5	1,7–3,3	3,3	2,6–4,1
45–54	6,5	5,8–7,3	5,7	4,6–6,8	7,3	6,2–8,4
55–64	13,5	12–15	12,1	9,7–14,4	14,8	12,9–16,7
≥ 65	19,8	18,2–21,4	18	15,2–20,7	21,2	19,1–23,4
Escolaridade (anos)						
Analfabeto/Fundamental incompleto	9,6	8,9–10,3	6,7	5,8–7,6	12,3	11,3–13,4
Fundamental completo/Médio incompleto	5,4	4,4–6,3	5,4	3,8–6,9	5,4	4,3–6,4
Médio completo/Superior incompleto	3,4	3–3,9	3,6	2,8–4,3	3,3	2,7–3,9
Superior completo	4,2	3,3–5	5,7	4–7,4	3,1	2,2–3,9
Raça/cor <sup>a</sup>						
Branco	6,7	6,1–7,2	6	5,2–6,8	7,3	6,5–8
Preto	7,2	5,8–8,5	5,4	3,2–7,6	8,7	7,1–10,4
Pardo	5,5	5,1–6	4,6	3,9–5,2	6,4	5,8–7
Categorias de IMC <sup>b</sup>						
Baixo peso/Normal (< 25 kg/m <sup>2</sup> )	3,3	2,8–3,8	3,4	2,7–4,2	3,2	2,5–3,8
Sobrepeso (entre 25 e 29,9 kg/m <sup>2</sup> )	6,9	6,1–7,7	6,5	5,3–7,6	7,5	6,4–8,6
Obesidade (≥ 30 kg/m <sup>2</sup> )	11,8	10,4–13,1	10,3	8,5–12,1	13	11,2–14,8

Estimativa populacional ("global")

Estimativa subpopulacional (setor sexo masculino)

Estimativa subpopulacional (setor sexo feminino)

Estimativa subpopulacionais (setor sexo masculino + escolaridade)

Estimativa subpopulacionais (setor sexo feminino + Raça/cor)

Estimativa subpopulacional (setor categorias de IMC)

# Estimação de proporções dentro de setores

- ▶ Nesses casos, a variável de pesquisa  $Y$  seria dada por:

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

- ▶ Na população como um todo, a proporção de unidades com atributo  $C$  é definida como  $P = A/N$  e a estimacão desta proporção foi discutida nas **aulas 17, 18 e 19**.



## Estimação de proporções dentro de setores

- Considere a notação a seguir. O número de unidades no setor  $j$  que também possuem o atributo  $C$  é definido como:

$$A_j = \sum_{k=1}^{n_j} Y_{ik}.$$

- E a proporção de unidades no setor  $j$  que também possuem o atributo  $C$  é definida como:

$$P_j = \frac{A_j}{N_j},$$

em que  $N_j$  é o tamanho do setor  $j$ .

## Estimação de proporções dentro de setores

- Sob **amostragem aleatória simples**, o estimador para  $P_j$  pode ser obtido a partir do estimador:

$$\hat{P}_j = p_j = \frac{1}{n_j} \sum_{k=1}^{n_j} Y_{ik} = \frac{a_j}{n_j}$$

em que  $a_j$  denota o **número de unidades na amostra no setor  $j$**  que também **possuem o atributo  $C$** .

## Estimação de proporções dentro de setores

- Considerando fixado o tamanho da amostra no setor  $j$ , a **variância condicional** do estimador  $p_j$  é dada por:

$$\text{Var}(\hat{p}_j) = \left(1 - \frac{n_j}{N_j}\right) \frac{P_j(1 - P_j)}{(n_j - 1)}.$$

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

$$\widehat{\text{Var}}(\hat{p}_j) = \left(1 - \frac{n_j}{N_j}\right) \frac{p_j(1 - p_j)}{(n_j - 1)}.$$

## Estimação de proporções dentro de setores

- ▶ Nas expressões acima,  $n_j$ ,  $N_j$  e  $P_j$  são, respectivamente, o número de unidades da amostra que pertencem ao setor  $j$ , o número total de unidades da população no setor  $j$  e a proporção de unidades no domínio que possuem o atributo  $C$ .

## Estimação de proporções dentro de setores

- Caso  $N_j$  não seja conhecido, a **fração de amostragem** no setor,  $n_j/N_j$ , pode ser aproximada por  $n/N$  na expressão anterior, levando ao estimador:

$$\widehat{\text{Var}}(p_j) = \left(1 - \frac{n}{N}\right) \frac{p_j q_j}{n_j - 1},$$

em que  $q_j = 1 - p_j$ .

## Estimação de proporções dentro de setores

- ▶ Para completar a inferência sobre uma proporção de unidades portadoras do atributo  $C$  no setor  $j$ , admite-se a validade da **aproximação normal** para a distribuição de  $p_j$  e soma-se uma **correção de continuidade**.
- ▶ Assim a expressão do **intervalo de confiança** para a proporção populacional  $p_j$  é dada por:

$$IC(P_j; 100 \times (1 - \alpha)\%) = \left[ p_j \pm \left( z_{\alpha/2} \sqrt{\widehat{\text{Var}}(p_j)} + \frac{1}{2n_j} \right) \right],$$

em que  $1/2n_j$  é a **correção de continuidade**.

- ▶ Essa correção é praticamente nula quando  $n_j$  cresce.

## Estimação de proporções dentro de setores

- ▶ Para se estimar o **número total**,  $A_j$ , de unidades da categoria  $C$  que estão no Setor  $j$ , há duas possibilidades.
- ▶ Se  $N_j$ , o número total de unidades da população que pertencem ao Setor  $j$ , for conhecido, pode-se usar a estimativa condicional:

$$\hat{A}_j = N_j \times \frac{a_j}{n_j} = N_j p_j.$$

- ▶ Se  $N_j$  não é conhecido, a estimativa é

$$\hat{A}_j = N \times \frac{a_j}{n}.$$

## Exemplo



## Exemplo

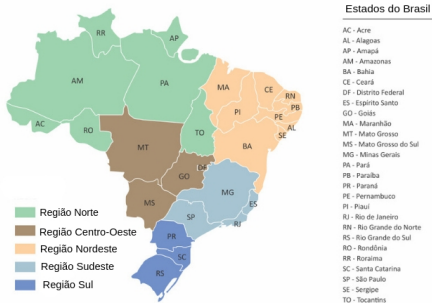
- ▶ Vamos estimar, a partir de uma **amostra aleatória simples sem reposição** com  $n = 300$ , a **proporção de municípios com população<sup>2</sup> menor que 10.000 habitantes** para cada **macro-região do Brasil**.

---

<sup>2</sup>**Atenção:** aqui temos como **população alvo** um conjunto de municípios; uma característica de interesse nos elementos (município) desta população é o **tamanho da população de habitantes**.

## Exemplo

- ▶ **População alvo:** municípios do Brasil.
- ▶ **Característica de interesse:** tamanho da população de habitantes < 10.000;
- ▶ **Setores:** macro-regiões do Brasil.



# Exemplo

```
# Dados dos municípios (população)
mun_aas <- readRDS(file = here::here("dados",
                                     "MunicBR_amostra.rds"))
```

```
mun_aas
```

```
## # A tibble: 300 x 9
```

	CodMunic	SiglaUF	CodUF	Pop	Area	Densidade	Regiao	Pop_menor_10	cpf
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<int>
## 1	1100031	RO	11	6495	1314.	4.94	Norte	1	5570
## 2	1100064	RO	11	19190	1451.	13.2	Norte	0	5570
## 3	1100346	RO	11	17399	3029.	5.74	Norte	0	5570
## 4	1100700	RO	11	13939	3442.	4.05	Norte	0	5570
## 5	1101435	RO	11	7883	807.	9.77	Norte	1	5570
## 6	1200179	AC	12	9836	1703.	5.78	Norte	1	5570
## 7	1200385	AC	12	17795	1943.	9.16	Norte	0	5570
## 8	1300631	AM	13	17332	17251.	1.00	Norte	0	5570
## 9	1301159	AM	13	26722	2631.	10.2	Norte	0	5570
## 10	1301852	AM	13	44503	2214.	20.1	Norte	0	5570

```
## # i 290 more rows
```

# Exemplo

```
# Estimativa populacional
round(mean(mun_aas$Pop_menor_10), 2)
```

```
## [1] 0.44
```

```
# Estimativa subpopulacionais
library(dplyr)
```

```
mun_aas %>%
  group_by(Regiao) %>%
  summarize(round(mean(Pop_menor_10), 2))
```

```
## # A tibble: 5 x 2
```

```
##   Regiao      `round(mean(Pop_menor_10), 2)`
```

```
##   <chr>                                <dbl>
```

```
## 1 Centro-Oeste                        0.71
```

```
## 2 Nordeste                            0.25
```

```
## 3 Norte                              0.36
```

```
## 4 Sudeste                            0.48
```

```
## 5 Sul                                0.62
```

## Exemplo

- ▶ O pacote `dplyr` nos ajudou na estimativa das proporções por setores (agrupadas).
- ▶ Vamos aproveitar este exemplo para apresentar o pacote `srvyr`, que utiliza uma sintaxe semelhante a utilizada pelo pacote `dplyr` na estimação de quantidades populacionais a partir de levantamentos por amostragem.
  - ▶ Veja a vinheta do pacote `srvyr`.

# Exemplo

```
# install.packages(srvyr)
library(srvyr)

mun_des <- mun_aas %>%
  as_survey_design(ids = 1,
                  fpc = cpf)

summary(mun_des)
```

```
## Independent Sampling design
## Called via srvyr
## Probabilities:
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05386 0.05386 0.05386 0.05386 0.05386 0.05386
## Population size (PSUs): 5570
## Data variables:
## [1] "CodMunic"      "SiglaUF"          "CodUF"            "Pop"              "Area"
## [8] "Pop_menor_10"  "cpf"
```

## Exemplo

```
# Estimativa populacional
mun_des %>%
  summarize(
    Proporção = survey_mean(Pop_menor_10, vartype = "ci")) %>%
  round(2)
```

```
## # A tibble: 1 x 3
##   Proporção Proporção_low Proporção_upp
##   <dbl>         <dbl>         <dbl>
## 1      0.44      0.39      0.49
```

## Exemplo

```
# Estimativas subpopulacionais
```

```
mun_des %>%
  group_by(Regiao, Pop_menor_10) %>%
  summarize(Proporção = survey_mean()) %>%
  mutate_at(vars(matches("Proporção")), function(x){round(x, 2)})
```

```
## # A tibble: 10 x 4
```

```
## # Groups:   Regiao [5]
```

##	Regiao	Pop_menor_10	Proporção	Proporção_se
##	<chr>	<dbl>	<dbl>	<dbl>
##	1 Centro-Oeste	0	0.29	0.09
##	2 Centro-Oeste	1	0.71	0.09
##	3 Nordeste	0	0.75	0.04
##	4 Nordeste	1	0.25	0.04
##	5 Norte	0	0.64	0.08
##	6 Norte	1	0.36	0.08
##	7 Sudeste	0	0.52	0.05
##	8 Sudeste	1	0.48	0.05
##	9 Sul	0	0.38	0.06
##	10 Sul	1	0.62	0.06



## Exemplo

```
# Estimativas subpopulacionais
mun_des %>%
  group_by(Regiao, Pop_menor_10) %>%
  summarize(Proporção = survey_mean(),
            a = unweighted(n())) %>%
  mutate_at(vars(matches("Proporção")), function(x){round(x, 2)})
```

```
## # A tibble: 10 x 5
## # Groups:   Regiao [5]
##   Regiao      Pop_menor_10 Proporção Proporção_se      a
##   <chr>          <dbl>      <dbl>      <dbl> <int>
## 1 Centro-Oeste      0      0.29      0.09      7
## 2 Centro-Oeste      1      0.71      0.09     17
## 3 Nordeste          0      0.75      0.04     67
## 4 Nordeste          1      0.25      0.04     22
## 5 Norte             0      0.64      0.08     21
## 6 Norte             1      0.36      0.08     12
## 7 Sudeste           0      0.52      0.05     53
## 8 Sudeste           1      0.48      0.05     48
## 9 Sul               0      0.38      0.06     20
## 10 Sul              1      0.62      0.06     33
```

## Exemplo

```
# Estimativas subpopulacionais
mun_des %>%
  group_by(Regiao, Pop_menor_10) %>%
  summarize(Proporção = survey_mean(vartype = "ci"),
            Total = survey_total(vartype = "ci")) %>%
  mutate_at(vars(matches("Proporção")), function(x){round(x, 2)}) %>%
  mutate_at(vars(matches("Total")), function(x){round(x)}) %>%
  filter(Pop_menor_10 == 1) %>% select(-Pop_menor_10) %>% knitr::kable()
```

Regiao	Proporção	Proporção_low	Proporção_upp	Total	Total_low	Total_upp
Centro-Oeste	0.71	0.53	0.89	316	173	458
Nordeste	0.25	0.16	0.33	408	248	569
Norte	0.36	0.20	0.52	223	102	344
Sudeste	0.48	0.38	0.57	891	665	1117
Sul	0.62	0.49	0.75	613	420	806

## Para casa

- ▶ Revisar os tópicos discutidos nesta aula.
- ▶ Faça uma busca por um levantamento por amostragem que tenha apresentado (parte dos) seus resultados em termos de proporções. Investigue os procedimentos de amostragem e métodos de estimação. Compartilhe os seus achados no Fórum Geral do Moodle.
- ▶ Estime a proporção (percentual) de municípios com população menor que 20.000 habitantes, com os seus respectivos erros padrões e intervalos de confiança de 95% (**dados no Moodle**).
  - ▶ A partir das estimativas pontuais, construa um **mapa das regiões do Brasil** para apresentar os resultados.
  - ▶ Estime para o **total** de municípios com **menos que 20.000 habitantes**.
- ▶ Compartilhe os seus achados no Fórum Geral do Moodle.

# Próxima aula

- ▶ **Área 3:** dimensionamento de amostra.

# Por hoje é só!

Bons estudos!

