MAT02025 - Amostragem 1

AAS: proporções das 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, 2021

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 **domínio** (setor, subgrupo ou subpopulação) 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 domínios da população: mulheres de 15 anos ou mais e homens de 18 anos ou mais.

Estimação de proporções dentro de setores

Rev Saude Publica. 2017;51 Supl 1:12s

Suplemento DCNT e Inquéritos
Artigo Original

RSP Revista de Saúde Pública

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

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

Deborah Carvalho Malta[I], Regina Tomie Ivata Bernal[II], Betine Pinto Moehlecke Iser[III,IV], Célia Landmann Szwarcwald[V], Bruce Bartholow Duncan[III], Maria Inês Schmidt[IV]

[I] Departamento de Enfermagem Materno Infantil e Saúde Pública. Escola de Enfermagem. Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brasil
[II] Núcleo de Pesquisas Epidemiológicas em Nutrição e Saúde. Universidade de São Paulo. São Paulo, SP, Brasil
[III] Programa de Pós-Graduação em Epidemiologia. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS, Brasil
[IV] Faculdade de Medicina. Universidade do Sul de Santa Catarina. Tubarão, SC, Brasil
[V] 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[a]						
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[b]						
Baixo peso/Normal (< 25 kg/m²)	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/m²)	6,9	6,1–7,7	6,5	5,3–7,6	7,5	6,4–8,6
Obesidade (≥ 30 kg/m²)	11,8	10,4–13,1	10,3	8,5–12,1	13	11,2–14,8

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 estimação 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 domínio 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 domínio j que também possuem o atributo C é definida como:

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

Estimação de proporções dentro de setores

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

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

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

Estimação de proporções dentro de setores

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

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

Estimação de proporções dentro de setores

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

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

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

- Essa correção é, praticamente, nula quando n_j cresce.

Exemplo

Exemplo

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

Exemplo

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

# Criando a variável Região
mun$Regiao <- NA

mun$Regiao[mun$SiglaUF %in% c("RS", "SC", "PR")] <- "Sul"
mun$Regiao[mun$SiglaUF %in% c("SP", "MG", "RJ", "ES")] <- "Sudeste"
mun$Regiao[mun$SiglaUF %in% c("MS", "MT", "GO", "DF")] <- "Centro-Oeste"
mun$Regiao[mun$SiglaUF %in% c("RO", "AC", "AM", "PA", "TO", "RR", "AP")] <- "No
mun$Regiao[mun$SiglaUF %in% c("BA", "SE", "AL", "PE", "PB", "PI", "MA", "CE", "

mun$Regiao <- factor(mun$Regiao)
```

Exemplo

```
# Sorteio da amostra
set.seed(2810)

cod_amostra <- sample(x = mun$CodMunic,
                      size = 300,
                      replace = F)

mun_amostra <- mun[which(mun$CodMunic %in% cod_amostra),]

# Criando a variável Pop < 10 mil hab.
mun_amostra$Pop_menor_10 <- ifelse(mun_amostra$Pop < 10000, 1, 0)

# cpf
mun_amostra$cpf <- length(mun$CodMunic)
```

Exemplo

```
mean(mun_amostra$Pop_menor_10)
```

```
## [1] 0.45
```

```
by(data = mun_amostra$Pop_menor_10,
   INDICES = mun_amostra$Regiao,
   FUN = mean)
```

```
## mun_amostra$Regiao: Centro-Oeste
## [1] 0.4333333
## ----------------------------------------------------------------------------
## mun_amostra$Regiao: Nordeste
## [1] 0.4285714
## ----------------------------------------------------------------------------
## mun_amostra$Regiao: Norte
## [1] 0.3913043
## ----------------------------------------------------------------------------
## mun_amostra$Regiao: Sudeste
## [1] 0.4631579
## ----------------------------------------------------------------------------
## mun_amostra$Regiao: Sul
## [1] 0.4918033
```

Exemplo

```r
library(survey)

mun_des <- svydesign(ids = ~1,
                     fpc = ~cpf,
                     data = mun_amostra)


svyby(formula = ~Pop_menor_10,
      by = ~Regiao,
      design = mun_des,
      FUN = svyciprop)
```

```
##                    Regiao Pop_menor_10 se.as.numeric(Pop_menor_10)
## Centro-Oeste Centro-Oeste    0.4333333                  0.08814892
## Nordeste         Nordeste    0.4285714                  0.05054458
## Norte               Norte    0.3913043                  0.09915081
## Sudeste           Sudeste    0.4631579                  0.04984582
## Sul                   Sul    0.4918033                  0.06236623
```

Para casa

- Revisar os tópicos discutidos nesta aula.
- 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%.
  - A partir das estimativas pontuais, construa um mapa das regiões do Brasil para apresentar os resultados.
  - Pense em um estimador 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!