MAT02025 - Amostragem 1

AAS: estimação de proporções para classificações em mais de duas categorias

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, 2022

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

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
1° ano	110	0,110
2° ano	108	0,108
3° ano	110	0,110
4° ano	115	0,115
5° ano	104	0,104
6° ano	119	0,119
7° ano	116	0,116
8° ano	107	0,107
9° ano	111	0,111
Total	**1.000**	**1,000**

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

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

- Gostaríamos de estimar $P_1, P_2, \ldots, P_9$, 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 $C \in \{1, 2, \ldots, m\}$, 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:

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

Proporções: duas ou mais categorias

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

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

Proporções: duas ou mais categorias

Obeservações

- 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$stype)) # 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
```

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.

Próxima aula

- Proporções e totais das subpopulações.

Por hoje é só!

Bons estudos!