

MAT02025 - Amostragem 1

AAS: estimativa de um índice

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



Estimativa de um índice

Relembrando

No sentido mais simples do termo, podemos dizer que um **número índice** é um quociente que expressa uma dada quantidade em comparação a uma **quantidade base**. Em outras palavras, são **valores relativos**.

Estimativa de um índice

- ▶ Frequentemente, a quantidade que deve ser estimada a partir de uma amostra aleatória simples é a **razão de duas variáveis**, ambas as quais variam de unidade para unidade.
- ▶ Em um levantamento por amostragem domiciliar, os exemplos são:
 - ▶ o **número de aparelhos de celular por residente**;
 - ▶ a **despesa com aplicativos de transporte por residente adulto**;
 - ▶ o **número médio de horas por semana gastas assistindo programas no serviço de *streaming* por criança de 10 a 15 anos**.

Estimativa de um índice

- ▶ A fim de estimar a primeira dessas quantidades, registraríamos para o i -ésimo domicílio ($i = 1, 2, \dots, n$) o número de residentes X_i que ali vivem e o número total de aparelhos de celular Y_i que eles possuem.
- ▶ O **parâmetro da população** a ser estimado é a **razão (ou índice)**

$$R = \frac{\text{número total de aparelhos de celular}}{\text{número total de residentes}} = \frac{\sum_{i=1}^N Y_i}{\sum_{i=1}^N X_i}.$$

Estimativa de um índice

- ▶ A estimativa amostral correspondente é

$$\hat{R} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} = \frac{\bar{y}}{\bar{x}}.$$

Estimativa de um índice

- ▶ Exemplos dessa natureza ocorrem, frequentemente, quando a unidade de amostragem (**no caso o domicílio**) compreende um grupo ou um conjunto de elementos (**residentes**) e nosso interesse está no valor médio da população **por elemento**.
- ▶ Os índices também aparecem em muitas outras aplicações, como, por exemplo
 - ▶ o índice de empréstimos para construções imobiliárias no total de empréstimos de um banco;
 - ▶ ou índice de acres plantados com trigo, no total de acres cultivados de uma fazenda;
 - ▶ ou índice de casos de diabetes não diagnosticado, no total de casos de diabetes.

Estimativa de um índice

Epidemiology/Health Services Research

ORIGINAL ARTICLE

Prevalence of Diabetes and High Risk for Diabetes Using A1C Criteria in the U.S. Population in 1988–2006

CATHERINE C. COWIE, PHD¹
KEITH F. RUST, PHD²
DANITA D. BYRD-HOLT, BBA³
EDWARD W. GREGG, PHD⁴

EARL S. FORD, MD⁵
LINDA S. GEISS, MS⁴
KATHLEEN E. BAINBRIDGE, PHD³
JUDITH E. FRADKIN, MD¹

cal trials in type 1 and type 2 diabetic patients, which have established widely accepted A1C treatment goals for diabetes. A cut point of $\geq 6.5\%$ for the diagnosis of diabetes was recommended by the

Prevalence of diabetes using A1C

Table 1—Crude prevalence of diagnosed diabetes, undiagnosed diabetes (A1C $\geq 6.5\%$), total diabetes (diagnosed and undiagnosed combined), total diabetes that is undiagnosed, and at high risk for diabetes (A1C ≥ 6.0 to $< 6.5\%$), by age, sex, and race/ethnicity: NHANES 2003–2006 (n = 13,094)

	Diagnosed diabetes	Undiagnosed diabetes	Total diabetes	Total diabetes that is undiagnosed	At high risk for diabetes
Combined age-groups (years)					
≥ 12	6.8 (6.1–7.5)	1.6 (1.2–1.9)	8.4 (7.6–9.2)	19.0 (15.2–22.7)	3.1 (2.7–3.4)
≥ 20	7.8 (7.0–8.6)	1.8 (1.4–2.2)	9.6 (8.7–10.5)	19.0 (15.2–22.7)	3.5 (3.0–3.9)
≥ 65	17.7 (15.6–19.7)	3.5 (2.6–4.3)	21.1 (18.7–23.5)	16.3 (12.9–19.8)	8.1 (6.6–9.6)

Estimativa de um índice

- ▶ A **distribuição amostral** de \hat{R} é mais complicada que a de \bar{y} , porque tanto o numerador \bar{y} , quanto o denominador, \bar{x} , variam de amostra para amostra.
- ▶ Em **pequenas amostras**, a distribuição de \hat{R} é **assimétrica**, e \hat{R} é, geralmente, uma estimativa ligeiramente **enviesada** de R .
- ▶ Em **grandes amostras**, a distribuição de \hat{R} tende à **normalidade** e o **viés** torna-se **insignificante**.

Estimativa de um índice

- ▶ O seguinte **resultado aproximado** servirá para a maioria dos propósitos¹.

Teorema

Se as variáveis Y_i e X_i são medidas em cada unidade de uma amostra aleatória simples de tamanho n , que se presume grande, a variância de $\hat{R} = \bar{y}/\bar{x}$ é, aproximadamente,

$$\text{Var}(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1},$$

em que $R = \bar{Y}/\bar{X}$ é o **índice dos valores médios da população**, e $f = n/N$ é a **fração de amostragem**.

¹A distribuição de \hat{R} é estudada com mais detalhes no Capítulo 6 de Cochran (1965) e no Capítulo 5 de Bolfarine e Bussab (2005).

Estimativa de um índice

Demonstração. Note que

$$\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}}.$$

- ▶ Se n é grande, \bar{x} não deve ser muito diferente de \bar{X} .
- ▶ Para evitar ter que calcular a distribuição da razão de duas variáveis aleatórias $\bar{y} - R\bar{x}$ e \bar{x} , substituímos \bar{x} por \bar{X} no denominador da expressão acima como uma aproximação. Isso dá

$$\hat{R} - R \doteq \frac{\bar{y} - R\bar{x}}{\bar{X}}.$$

Estimativa de um índice

- ▶ Agora calcule a média de todas as amostras aleatórias simples de tamanho n :

$$E(\hat{R} - R) = \frac{E(\bar{y} - R\bar{x})}{\bar{X}} = \frac{\bar{Y} - R\bar{X}}{\bar{X}},$$

uma vez que $R = \bar{Y}/\bar{X}$.

- ▶ Isso mostra que, para a ordem de aproximação usada aqui, \hat{R} é uma estimativa não enviesada de R .

Estimativa de um índice

- ▶ Da expressão aproximada, também obtemos

$$\text{Var}(\hat{R}) = E(\hat{R} - R)^2 \doteq \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2.$$

- ▶ A quantidade $\bar{y} - R\bar{x}$ é a média amostral da variável $D_i = Y_i - RX_i$, cuja média populacional, $\bar{D} = \bar{Y} - R\bar{X}$, é igual a 0.
- ▶ Portanto, podemos encontrar $\text{Var}(\hat{R})$ aplicando o teorema para a variância da média de uma amostra aleatória simples à variável D_i e dividindo por \bar{X}^2 .

Estimativa de um índice

► Isso dá

$$\begin{aligned}\text{Var}(\hat{R}) &\doteq \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2 = \frac{1}{\bar{X}^2} \frac{S_D^2}{n} (1 - f) \\ &= \frac{1 - f}{n\bar{X}^2} \frac{\sum_{i=1}^N (D_i - \bar{D})^2}{N - 1} = \frac{1 - f}{n\bar{X}^2} \frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N - 1},\end{aligned}$$

o que **completa a demonstração**.

Estimativa de um índice

- ▶ Como estimativa amostral de

$$\frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N - 1}$$

é comum tomarmos

$$\frac{\sum_{i=1}^n (Y_i - \hat{R}X_i)^2}{n - 1}.$$

- ▶ Pode-se demonstrar que essa estimativa tem um **viés de ordem $1/n$** .²

²Ou seja, conforme $n \rightarrow \infty$ o viés decresce a zero mais rapidamente que a sequência $1/n$ ($\text{Viés}_n = o(n^{-1})$).

Estimativa de um índice

- Para o **erro padrão estimado** de \hat{R} , temos

$$s_{\hat{R}} = \frac{\sqrt{1-f}}{\sqrt{n}\bar{X}} \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{R}X_i)^2}{n-1}}.$$

- Se \bar{X} não é conhecido, a estimativa amostral \bar{x} o substitui no denominador da fórmula.
- Uma fórmula prática para calcular $s_{\hat{R}}$ é dada por

$$s_{\hat{R}} = \frac{\sqrt{1-f}}{\sqrt{n}\bar{X}} \sqrt{\frac{\sum_{i=1}^n Y_i^2 - 2\hat{R} \sum_{i=1}^n Y_i X_i + \hat{R}^2 \sum_{i=1}^n X_i^2}{n-1}}.$$

Para casa

Para casa

- ▶ A tabela a seguir mostra o número de pessoas (X_1), a renda familiar semanal (X_2) e os gastos semanais com alimentação (Y) em uma amostra aleatória simples de 33 famílias de baixa renda.
- ▶ Como a amostra é pequena, os dados se destinam apenas a ilustrar os cálculos.

Para casa

Table 1: Renda semanal e custo dos alimentos de 33 famílias

Número da família	Tamanho (x_1)	Renda (x_2)	Custo dos alimentos (y)	Número da família	Tamanho (x_1)	Renda (x_2)	Custo dos alimentos (y)
1	2	62	14,3	18	4	83	36,0
2	3	62	20,8	19	2	85	20,6
3	3	87	22,7	20	4	73	27,7
4	5	65	30,5	21	2	66	25,9
5	4	58	41,2	22	5	58	23,3
6	7	92	28,2	23	3	77	39,8
7	2	88	24,2	24	4	69	16,8
8	4	79	30,0	25	7	65	37,8
9	2	83	24,2	26	3	77	34,8
10	5	62	44,4	27	3	69	28,7
11	3	63	13,4	28	6	95	63,0
12	6	62	19,8	29	2	77	19,5
13	4	60	29,4	30	2	69	21,6
14	4	75	27,1	31	6	69	18,2
15	2	90	22,2	32	4	67	20,1
16	5	15	37,7	33	2	63	20,7
17	3	69	22,6				

Para casa

1. Estime a partir da amostra
 - a. o gasto semanal médio com comida por família;
 - b. o gasto semanal médio com comida por pessoa;
 - c. a porcentagem da renda que é gasta com comida.
2. Calcule os erros padrões dessas estimativas (pode ignorar as **cpf**).
3. Compartilhe os seus resultados no Fórum Geral do Moodle.

Próxima aula

- ▶ Estimativa da média e do total em subpopulações.

Por hoje é só!

Bons estudos!

