MAT02025 - Amostragem 1

AAS: estimativa de valores médios e totais 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, 2023

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

Estimativa do valor médio das subpopulações

Estimativa do valor médio das subpopulações

- Em muitos levantamentos, as estimativas são feitas para cada uma das várias **classes** **(setores ou domínios)** nas quais a população está subdividida.
- Em um levantamento domiciliar, estimativas separadas podem ser necessárias para **famílias de** $0, 1, 2, \ldots$ **filhos**, para **proprietários** e **locatários**, ou para famílias em diferentes **grupos de ocupação**.

Estimativa do valor médio das subpopulações

Table 2

Prevalence of diabetes mellitus in adults overall and by sex according to sociodemographic and clinical characteristics. *Brazilian National Health Survey*, 2013 and 2019.

Characteristic	2013						2019					
	Total		Men		Women		Total		Men		Women	
	%	95%CI	%	95%CI	%	95%CI	%	95%CI	%	95%CI	%	95%CI
Total	6.2	5.9-6.6	5.3	4.8-5.8	7.0	6.5-7.5	7.7	7.4-8.0	6.9	6.5-7.4	8.4	8.0-8.8
Age (years)												
18-24	0.5	0.3-0.8	0.4	0.1-0.7	0.6	0.2-1.1	0.7	0.4-1.1	1.0	0.4-1.7	0.4	0.2-0.6
25-34	0.8	0.6-1.0	0.7	0.4-1.1	0.9	0.6-1.2	0.9	0.7-1.2	0.7	0.3-1.0	1.1	0.8-1.5
35-44	2.9	2.4-3.4	2.4	1.7-3.2	3.4	2.6-4.1	3.1	2.7-3.6	3.1	2.5-3.8	3.2	2.5-3.8
45-54	6.6	5.8-7.4	5.8	4.6-6.9	7.3	6.2-8.4	7.7	6.9-8.5	6.9	5.8-8.0	8.4	7.2-9.6
55-64	13.5	12.1-14.9	12.1	9.8-14.3	14.7	12.8-16.6	14.8	13.8-15.7	13.6	12.2-15.0	15.8	14.5-17.1
≥ 65	19.8	18.2-21.3	17.7	15.1-20.3	21.4	19.3-23.5	21.6	20.5-22.7	20.2	18.6-21.9	22.7	21.2-24.1
Race/Skin color												
White	6.7	6.1-7.2	6.0	5.2-6.8	7.3	6.5-8.0	8.0	7.6-8.5	7.9	7.1-8.6	8.2	7.5-8.9
Black	7.3	6.0-8.6	5.5	3.4-7.5	8.9	7.2-10.5	7.8	7.0-8.7	6.9	5.7-8.0	8.7	7.5-9.9
Mixed-race	5.5	5.0-5.9	4.5	3.8-5.1	6.4	5.7-7.0	7.3	6.9-7.7	5.9	5.3-6.4	8.5	7.9-9.1
Asian	6.3	3.0-9.6	7.3	0.9-13.7	5.6	2.2-9.0	12.8	7.9-17.7	13.6	6.3-21.0	12.0	5.5-18.5
Indigenous	6.9	2.7-11.1	5.4	1.8-18.7 *	8.0	2.8-13.2	7.5	4.4-10.6	5.2	1.4-8.9	10.2	5.3-15.1
Education level												
Incomplete elementary	9.6	9.0-10.3	6.7	5.8-7.5	12.4	11.3-13.4	12.9	12.3-13.5	10.2	9.4-11.0	15.4	14.4-16.3
Complete elementary	5.4	4.5-6.3	5.4	4.0-6.9	5.4	4.3-6.5	6.3	5.5-7.0	5.4	4.4-6.4	7.1	6.1-8.2
Complete high school	3.4	2.9-3.8	3.5	2.8-4.2	3.3	2.7-3.8	4.6	4.2-5.0	4.6	4.0-5.2	4.6	4.1-5.1
Complete higher education	4.1	3.3-5.0	5.5	3.9-7.2	3.1	2.3-4.0	4.7	4.1-5.2	6.0	5.0-7.1	3.6	3.1-4.2

Fonte: http://dx.doi.org/10.1590/0102-311X00149321.

Estimativa do valor médio das subpopulações

- Na situação mais simples, cada unidade da população cai em um dos setors.
- Assuma que o j-ésimo setor contém N_j unidades e seja n_j o número de unidades em uma amostra aleatória simples de tamanho n que por acaso caem neste setor.
- Se $Y_{jk}(k = 1, 2, \ldots, n_j)$ são as medidas nessas unidades, a média da população $\overline{Y}_j$ para o j-ésimo setor é estimada por

$$\overline{y}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} Y_{jk}.$$

Estimativa do valor médio das subpopulações

- À primeira vista, $\overline{y}_j$ parece ser uma **estimativa de razão (índice)**.
- Embora n seja fixo, n_j variará de uma amostra de tamanho n para outra.
- A complicação de uma estimativa de razão pode ser evitada considerando a distribuição de $\overline{y}_j$ sobre as amostras nas quais n e n_j **são fixos**.
  - Assumimos $n_j > 0$.

Estimativa do valor médio das subpopulações

- Na totalidade das amostras, com n e n_j determinados, a probabilidade de que qualquer conjunto específico de n_j unidades das N_j unidades no setor j sejam sorteadas é

$$\frac{{}_{N-N_j}C_{n-n_j}}{{}_{N-N_j}C_{n-n_j} \times {}_{N_j}C_{n_j}} = \frac{1}{{}_{N_j}C_{n_j}}.$$

- Uma vez que cada conjunto específico de n_j unidades do setor j pode aparecer em todas as seleções de $(n - n_j)$ unidades, dentre as $(N - N_j)$ que não estão no setor j, o numerador acima é o número de amostras contendo um conjunto especificado de n_j e o denominador é o número total de amostras.

Estimativa do valor médio das subpopulações

- Segue-se que os **teoremas das aulas 9, 10 e 11** se aplicam ao Y_{jk} se colocarmos n_j no lugar de n e N_j no lugar de N.

Do **Teorema 9.1**, $\overline{y}_j$ é um estimador **não enviesado** para $\overline{Y}_j$.

Do **Teorema 10.1**, o **erro padrão** de $\overline{y}_j$ é $\frac{S_j}{\sqrt{n_j}}\sqrt{1-(n_j/N_j)}$, em que

$$S_j^2 = \frac{1}{N_j - 1}\sum_{k=1}^{N_j}(Y_{jk} - \overline{Y}_j)^2.$$

Estimativa do valor médio das subpopulações

De acordo com o **Teorema 11.1** e o **Corolário 11.1** , uma estimativa do erro padrão de $\overline{y}_j$ é

$$\frac{s_j}{\sqrt{n_j}}\sqrt{1-(n_j/N_j)},$$

em que

$$s_j^2 = \frac{1}{n_j - 1}\sum_{k=1}^{n_j}(Y_{jk} - \overline{y}_j)^2.$$

- Se o valor de N_j **não for conhecido**, a quantidade n/N pode ser utilizada em lugar de n_j/N_j, no cálculo das **cpf**.
  - Na amostragem aleatória simples, n_j/N_j é uma estimativa não enviesada de n/N.

Estimativa dos valores totais das subpopulações

Estimativa dos valores totais das subpopulações

- Suponha que em uma população (de adultos), na qual algumas pessoas são obesas (IMC > 30) e outras não, podemos desejar estimar, por meio de uma amostra, o ***total* de pessoas com diabetes *entre os obesos***.
- Se N_j (o **número de pessoas obesas** na população) é conhecido, não há problema.
  - A estimativa a partir da amostra é $N_j \overline{y}_j$ e seu erro padrão condicional é N_j vezes $\frac{S_j}{\sqrt{n_j}} \sqrt{1 - (n_j/N_j)}$.

Estimativa dos valores totais das subpopulações

- Alternativamente, se o *total de indivíduos com diabetes* for conhecido na **população**, uma estimativa de razão pode ser usada.
  - A **amostra** fornece uma estimativa da razão

$$\frac{\text{total de pessoas com diabetes entre os obesos}}{\text{total de indivíduos com diabetes}}.$$

- Isso é multiplicado pelo *total de indivíduos com diabetes* conhecido na **população**.

Estimativa dos valores totais das subpopulações

- Se nem N_j, e nem o *total de indivíduos com diabetes* é conhecido, essas estimativas não podem ser feitas.
- Em vez disso, multiplicamos o *valor amostral total* das unidades Y contidas no j-ésimo setor pelo **fator de expansão** N/n.
- Isso dá a estimativa

$$\hat{Y}_{T_j} = \frac{N}{n} \sum_{k=1}^{n_j} Y_{jk}.$$

- Mostraremos que $\hat{Y}_{T_j}$ **é imparcial** e obteremos seu erro padrão sobre amostras repetidas de tamanho n.
  - O artifício de manter n_j constante, bem como n não ajuda neste caso.

Estimativa dos valores totais das subpopulações

- Ao fazermos a demonstração, voltamos à notação original, na qual Y_i é a medida da i-ésima unidade da população.
- Defina para cada unidade na população uma nova variável Y_i', em que

$$Y_i' = \begin{cases} Y_i, & \text{se a unidade pertencer ao setor } j, \\ 0, & \text{caso contrário.} \end{cases}$$

- O valor **total populacional** da variável Y_i' é

$$\sum_{i=1}^{N} Y_i' = \sum_{\text{setor } j} Y_i = Y_{T_j}.$$

Estimativa dos valores totais das subpopulações

- Em uma **amostra aleatória simples** de tamanho n, $Y_i' = Y_i$ para todas as n_j unidades que se encontram no j-ésimo setor; $Y_i' = 0$ para todas as restantes $n - n_j$ unidades.
- Se $\overline{y}'$ é a média amostral de Y_i', então temos

$$N\overline{y}' = \frac{N}{n}\sum_{i=1}^{n} Y_i' = \frac{N}{n}\sum_{k=1}^{n_j} Y_{jk} = \hat{Y}_{T_j}$$

- Este resultado mostra que a estimativa $\hat{Y}_{T_j}$ é N vezes a média amostral de Y_i'.

Estimativa dos valores totais das subpopulações

- Em repetidas amostras de tamanho n, podemos aplicar os **teoremas das aulas 9, 10 e 11** às variáveis Y_i'.
- Estes, por sua vez, mostram que $\hat{Y}_{T_j}$ é uma **estimativa imparcial** de Y_{T_j} com erro padrão

$$\sigma(\hat{Y}_{T_j}) = \frac{NS'}{\sqrt{n}}\sqrt{1-(n/N)},$$

em que S' é desvio padrão populacional de Y_i'.

Estimativa dos valores totais das subpopulações

- Para calcular S' , consideramos a população como consistindo de N_j valores Y_i que estão no j-ésimo setor e de $N - N_j$ **valores zero**. Assim

$$S'^2 = \frac{1}{N-1}\left(\sum_{\text{setor } j} Y_i^2 - \frac{Y_{T_j}^2}{N}\right).$$

Estimativa dos valores totais das subpopulações

- Pelo teorema da **aula 11**, uma estimativa amostral do erro padrão de $\hat{Y}_{T_j}$ será

$$s(\hat{Y}_{T_j}) = \frac{Ns'}{\sqrt{n}}\sqrt{1-(n/N)}.$$

- No cálculo de s', qualquer unidade que não esteja no j-ésimo setor recebe um valor zero.

Comparação da eficiência dos estimadores de total no setor

- Às vezes é possível, com algum esforço, identificar e contar as unidades que não contribuem com nada, de modo que em nossa notação $(N - N_j)$, e portanto N_j, seja conhecido.
- Consequentemente, vale a pena examinar o quanto da $\mathrm{Var}\,(\hat{Y}_{T_j})$ é reduzido quando N_j é conhecido.
- Se N_j **não for conhecido**, temos (pelo **Corolário 10.2**)

$$\mathrm{Var}\,(\hat{Y}_{T_j}) = \frac{N^2 S'^2}{n}\left(1 - \frac{n}{N}\right).$$

Comparação da eficiência dos estimadores de total no setor

- Se $\overline{Y}_j$ e S_j são a média e o desvio padrão no setor de interesse (ou seja, entre as unidades diferentes de zero), é possível verificar que

$$(N-1)S'^2 = (N_j-1)S_j^2 + N_j\overline{Y}_j^2\left(1-\frac{N_j}{N}\right).$$

- Uma vez que os termos em $1/N_j$ e $1/N$ são quase sempre insignificantes, temos

$$S'^2 \doteq P_jS_j^2 + P_jQ_j\overline{Y}_j^2,$$

em que $P_j = N_j/N$ e $Q_j = 1 - P_j$.

Comparação da eficiência dos estimadores de total no setor

- Desta forma,

$$\text{Var}\left(\hat{Y}_{T_j}\right) \doteq \frac{N^2}{n}\left(P_j S_j^2 + P_j Q_j \overline{Y}_j^2\right)\left(1 - \frac{n}{N}\right). \tag{1}$$

Comparação da eficiência dos estimadores de total no setor

- Se as unidades diferentes de zero forem identificadas (ou seja, se N_j **for conhecido**), retiramos delas uma amostra de tamanho n_j. A estimativa do total do setor é $N_j\overline{y}_j$ com variância

$$\text{Var}\left(N_j\overline{y}_j\right) = \frac{N_j^2}{n_j}S_j^2\left(1-\frac{n_j}{N_j}\right) = \frac{N^2}{n_j}P_j^2S_j^2\left(1-\frac{n_j}{N_j}\right). \qquad (2)$$

Comparação da eficiência dos estimadores de total no setor

- As variâncias dadas pelas expressões (1) e (2) são comparáveis.
- Em (1), o número médio de unidades diferentes de zero na amostra de tamanho n é nP_j.
- Se tomarmos $n_j = nP_j$ em (2), de modo que o número de valores diferentes de zero a serem medidos seja aproximadamente o mesmo com ambos os métodos, (2) torna-se

$$\text{Var}\left(N_j \overline{y}_j\right) = \frac{N^2}{n} P_j S_j^2 \left(1 - \frac{n}{N}\right). \tag{3}$$

Comparação da eficiência dos estimadores de total no setor

- A razão entre as variâncias (3) e (1) é

$$\frac{\text{Var}_{N_j \text{ conhecido}}(N_j \overline{y}_j)}{\text{Var}_{N_j \text{ desconhecido}}(\hat{Y}_{T_j})} = \frac{S_j^2}{S_j^2 + Q_j \overline{Y}_j^2} = \frac{C_j^2}{C_j^2 + Q_j} \leq 1,$$

em que $C_j = S_j / \overline{Y}_j$ é o **coeficiente de variação** entre as unidades de valores diferentes de zero.

Comparação da eficiência dos estimadores de total no setor

Observação

- Como era de se esperar, a redução da variância, decorrente do conhecimento de N_j, é maior quando a proporção de unidades de valor nulo é grande e quando Y_j varia relativamente pouco entre as unidades de valor diferente de zero.

Para casa (PQP)

Para casa (PQP)

- Considere o exemplo da aula 11 (assinaturas de uma petição). Depois de selecionada a amostra, o número de folhas completamente cheias (com 42 assinaturas cada) foi contado e verificou-se que eram 326. Use essa informação para fazer uma estimativa melhorada do número total de assinaturas e achar o erro padrão da sua estimativa.

Próxima aula

- Validade da aproximação normal.

Por hoje é só!

Bons estudos!