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, 2021

Estimativa do valor médio das subpopulações

Estimativa do valor médio das subpopulações

- Em muitoss 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

- Na situação mais simples, cada unidade da população cai em um dos domínios.
- Assuma que o j-ésimo domínio conter N_j unidades e seja n_j o número de unidades em uma amostra aleatória simples de tamanho n que por acaso cai neste domínio.
- 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 domínio 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 (como na última aula).
- 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 domínio j sejam sorteadas é

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

- Uma vez que cada conjunto específico de n_j unidades do domínio j pode aparecer em todas as seleções de $(n - n_j)$ unidades, dentre as $(N - N_j)$ que não estão no domínio 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 para n e N_j para N.

Do teorema da aula 9, $\overline{y}_j$ é um estimador não enviesado para $\overline{Y}_j$.

Do teorema da aula 10, 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 da aula 11, 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

- Na relação de contas a receber de uma empresa, na qual algumas contas foram pagas e outras não, podemos desejar estimar por uma amostra o valor total (em reais) das contas não pagas.
- Se N_j (o número de contas não pagas na população) é conhecido, não há problema.
  - A estimativa 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 valor total a receber, de acordo com a relação das contas, for conhecido, uma estimativa de razão pode ser usada.
- A amostra fornece uma estimativa da razão (montante total de faturas não pagas) / (montante total de todas as faturas).
- Isso é multiplicado pelo valor total a receber conhecido na relação das contas.

Estimativa dos valores totais das subpopulações

- Se nem N_j nem o total a receber é conhecido, essas estimativas não podem ser feitas.
- Em vez disso, multiplicamos o valor amostral total das unidades Y contidas no j-ésimo domínio 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

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

- 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 domínio; $Y_i' = 0$ 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 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 domínio 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 domínio recebe um valor zero.

Comparação da eficiência dos estimadores de total no domínio

- 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 **Var** $(\hat{Y}_{T_j})$ é reduzido quando N_j é conhecido.
- Se N_j não for conhecido, temos

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

Comparação da eficiência dos estimadores de total no domínio

- Se $\overline{Y}_j$ e S_j são a média e o desvio padrão no domínio 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 domínio

- Desta forma,

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

- Se as unidades diferentes de zero forem identificadas, retiramos delas uma amostra de tamanho n_j. A estimativa do total do domínio é $N_j \overline{y}_j$ com variância

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

Comparação da eficiência dos estimadores de total no domínio

- As variância 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

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

Comparação da eficiência dos estimadores de total no domínio

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

$$\frac{\mathrm{Var}\,(N_j \text{ conhecido})}{\mathrm{Var}\,(N_j \text{ desconhecido})} = \frac{S_j^2}{S_j^2 + Q_j \overline{Y}_j^2} = \frac{C_j^2}{C_j^2 + Q_j},$$

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

Comparação da eficiência dos estimadores de total no domínio

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.

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.

```r
library(survey)
data(api)

# View(apisrs)
# ?apisrs

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

# Estimativa do total de alunos
# matriculados
svytotal(x = ~enroll, design = api.des)
```

Índice de Desempenho Acadêmico

```
##           total     SE
## enroll 3621074 169520
```

```
# Estimativa da média de alunos
# matriculados por escola
svymean(x = ~enroll, design = api.des)
```

```
##          mean     SE
## enroll 584.61 27.368
```

Índice de Desempenho Acadêmico

```
# Estimativa do total de alunos
# matriculados por tipo de escola
svyby(formula = ~enroll, by = ~stype, design = api.des, FUN = svytotal)
```

```
##   stype    enroll        se
## E     E 1849900.0  99738.62
## H     H  890666.2 187717.67
## M     M  880508.1 151805.23
```

```
# Estimativa da média de alunos
# matriculados por tipo de escola
svyby(formula = ~enroll, by = ~stype, design = api.des, FUN = svymean)
```

```
##   stype    enroll        se
## E     E  420.6479  12.76345
## H     H 1150.3600 117.20190
## M     M  861.5455  61.61652
```

Próxima aula

- Validade da aproximação normal.

Por hoje é só!

Bons estudos!