

MAT02036 - Amostragem 2

Aula 21 - Amostragem por Conglomerados em 2 Estágios - Introdução

Markus Stein

Departamento de Estatística, IME/UFRGS

2022/2

Housekeeping

- Aproveitem o momento presencial para tirar dúvidas
- Se estivéssemos no ensino remoto ou à distância
 - vocês poderiam estar somente ouvindo, sem interação
 - ou assistindo vídeos e material em outro momento
- Depois das aulas, rever material da aula passada
 - fazer exercícios
 - se preparar para a próxima aula

Aula passada

Estimação na AS

- O estimador tipo *Horvitz-Thompson* do total $T = \sum_{i=1}^K T_i$ sob AS,
 - o peso amostral das unidades da amostra é sempre igual a $d_i = 1/\pi_i = K$, então

$$\hat{T}_{AS} = K t_r = K \sum_{i \in s_r} y_i$$

em que $t_r = \sum_{i \in s_r} y_i$ é a soma amostral dos valores observados da variável y .

- Para estimar a média populacional $\bar{Y} = \frac{T}{N} = \frac{\sum_{r=1}^K t_r}{\sum_{r=1}^K n_r}$ um estimador não viciado (?) é dado por (quando N é conhecido)

$$\bar{y}_{AS} = \frac{\hat{T}_{AS}}{N} = \frac{K t_r}{N}.$$

Aula passada

Estimação na AS

Estimador	Observação
$\hat{T}_{AS} = K t_r = K \sum_{r=1}^K I(r) t_r$	
$\bar{y}_{AS} = \frac{K}{N} t_r$	se N é conhecido
$\bar{y}_{AS} = \frac{t_r}{n_r} = \bar{y}$	se N é desconhecido
$\widehat{Var}_{1AS}(\hat{T}_{AS}) = N^2 \widehat{Var}_{1AS}(\bar{y}_{AS})$	se N é conhecido e sem ordenação
$\widehat{Var}_{2AS}(\hat{T}_{AS}) = N^2 \widehat{Var}_{2AS}(\bar{y}_{AS})$	se N é conhecido e houver ordenação
$\widehat{Var}_{1AS}(\bar{y}_{AS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{i \in s_r} (y_i - \bar{y}_{AS})^2$	se N é conhecido e sem ordenação
$\widehat{Var}_{2AS}(\bar{y}_{AS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{2(n-1)} \sum_{i \in s_r} (y_i - y_{i+K})^2$	se N é conhecido e houver ordenação

Aula passada

Estimação na AS

Exemplo: .

Considere a população abaixo e $n = 2$:

$$X = (2, 6, 10, 8, 10, 12)$$

- a. Calcule $E(\bar{y}_{AS})$ e $Var(\bar{y}_{AS})$.
- b. Calcule $E\left[\widehat{Var}_{AS}(\bar{y}_{AS})\right]$.
- c. \bar{y}_{AS} e $\widehat{Var}_{AS}(\bar{y}_{AS})$ são **ENV** para os respectivos parâmetros a que se destinam estimar?

Amostragem conglomerada em dois estágios

Amostragem conglomerada em dois estágios

- *Amostragem Conglomerada em dois estágios* - **AC2** = **AC1** + subamostragem
 - **Estágio 1**: selecione uma amostra a de m UPAs (conglomerados).
 - **Estágio 2**: para cada UPA C_i tal que $i \in a$, selecione amostra s_i de n_i unidades secundárias das N_i unidades existentes nessa UPA.
- A **amostra completa** de unidades selecionadas é dada por:

$$s = s_{i_1} \cup s_{i_2} \cup \dots \cup s_{i_m} = \bigcup_{k=1}^m s_{i_k}$$

- O **tamanho total da amostra** é $n = \sum_{i \in a} n_i$.

Amostragem conglomerada em dois estágios

As principais razões para adotar amostragem conglomerada em dois estágios são as seguintes:

- 1) Geralmente não é prático pesquisar todas as unidades nos conglomerados selecionados: *conglomerados muito grandes, carga de trabalho variável por entrevistador, etc.*
- 2) Constatou-se que a perda de precisão da AC1S em relação à AAS para amostras de mesmo tamanho é tanto maior quanto maior for o tamanho do conglomerado. A adoção de AC2 vem reduzir a influência do tamanho dos conglomerados na eficiência da amostragem conglomerada, permitindo ao amostrista controlar melhor a precisão do estimador resultante, ao manejar o número de unidades que são selecionadas em cada conglomerado. Na AC1 isso não está sob controle do amostrista, pois uma vez selecionado um conglomerado, todas as suas unidades devem ser pesquisadas.
- 3) Se a variância dentro dos conglomerados for pequena, as médias por conglomerados \bar{Y}_i podem ser bem estimadas por amostragem.
- 4) *Amostragem em dois estágios é mais complexa, porém mais flexível.*

Amostragem conglomerada em dois estágios

- Na estimação sob AC2, cujo plano amostral compreende *dois estágios de seleção*, para encontrar médias e variâncias de estimadores, devem ser calculadas médias considerando todas as possíveis amostras sob o plano amostral.
 - Isto requer considerar todas as possíveis amostras no primeiro estágio e, todas as possíveis amostras no segundo estágio, dentro de cada amostra de UPAs do primeiro estágio.
- Usaremos os resultados de @Cochran1977, Expressão 10.1 da página 275 e Expressão 10.2 da página 276, respectivamente:

$$E[\hat{\theta}] = E_1[E_2(\hat{\theta})]$$

$$Var[\hat{\theta}] = Var_1[E_2(\hat{\theta})] + E_1[Var_2(\hat{\theta})]$$

- E_2 , V_2 denotam, respectivamente, valor esperado e variância considerando todas as possíveis amostras de unidades dentro de um conjunto fixado de UPAs (estágio 2).
- E_1 , V_1 denotam, respectivamente, valor esperado e variância considerando todas as possíveis amostras de UPAs (estágio 1).

Amostragem conglomerada em dois estágios

- O estimador não viciado de Horvitz-Thompson do total T sob AC2 é dado por:

$$\hat{T}_{AC2/HT} = \sum_{i \in a} \frac{\hat{Y}_i}{\pi_i} = \sum_{i \in a} \frac{1}{\pi_i} \sum_{j \in s_i} \frac{y_{ij}}{\pi_{j|i}} = \sum_{i \in a} \sum_{j \in s_i} d_{ij} y_{ij}$$

onde:

- π_i é a probabilidade de inclusão da UPA i ;
- s_i é a amostra de unidades selecionadas dentro da UPA i ;
- $\hat{T}_i = \sum_{j \in s_i} \frac{y_{ij}}{\pi_{j|i}}$ é um estimador HT do total Y_i da UPA i ;
- $\pi_{j|i} = P(j \in s_i \mid i \in a)$ é a probabilidade de inclusão da unidade j dado que a UPA i está na amostra a ; e
- $d_{ij} = \pi_{ij}^{-1} = \pi_i^{-1} \pi_{j|i}^{-1}$ é o peso associado à unidade j da UPA i .

Amostragem conglomerada em dois estágios

- A variância de $\hat{Y}_{AC2/HT}$ sob AC2 é dada por:

$$\begin{aligned} Var_{AC2} \left(\hat{T}_{AC2/HT} \right) &= Var_1 \left[E_2 \left(\sum_{i \in a} \frac{\hat{T}_i}{\pi_i} \right) \right] + E_1 \left[Var_2 \left(\sum_{i \in a} \frac{\hat{T}_i}{\pi_i} \right) \right] \\ &= Var_1 \left[\sum_{i \in U} R_i E_2 \left(\hat{T}_i \right) / \pi_i \right] + E_1 \left[\sum_{i \in U} R_i Var_2 \left(\hat{T}_i \right) / \pi_i^2 \right] \\ &= Var_1 \left(\sum_{i \in a} T_i / \pi_i \right) + \sum_{i \in U} Var_2 \left(\hat{T}_i \right) / \pi_i \\ &= Var_{UPA} + Var_{USA} \end{aligned}$$

onde:

- R_i é a variável indicadora da presença da unidade i na amostra;
- Var_{UPA} é a componente de variância de $\hat{T}_{AC2/HT}$ proveniente da amostragem de UPAs (estágio 1), isto é, variância caso **AC1S** fosse usada (sem subamostragem); e
- Var_{USA} é a componente de variância de $\hat{T}_{AC2/HT}$ proveniente da amostragem de USAs (amostragem no estágio 2).

Amostragem conglomerada em dois estágios

- Um estimador não viciado (de Horvitz-Thompson) da média por unidade \bar{Y} é dado por:

$$\bar{y}_{AC2/HT} = \frac{\hat{T}_{AC2/HT}}{N} = \frac{1}{N} \left(\sum_{i \in a} \frac{\hat{T}_i}{\pi_i} \right)$$

- Se N for conhecido, um estimador tipo razão para estimar o total T é dado por:

$$\hat{Y}_{AC2}^R = N \left(\sum_{i \in a} \frac{\hat{T}_i}{\pi_i} \right) / \left(\sum_{i \in a} \frac{N_i}{\pi_i} \right)$$

Um estimador tipo razão da média por unidade é dado por:

$$\bar{y}_{AC2}^R = \left(\sum_{i \in a} \frac{\hat{T}_i}{\pi_i} \right) / \left(\sum_{i \in a} \frac{N_i}{\pi_i} \right)$$

Este estimador de razão da média pode ser calculado mesmo quando N for desconhecido.

Amostragem conglomerada em dois estágios

AC2 com AAS nos 2 estágios

- Tratamos agora do plano amostral *AC2 com AAS nos 2 estágios* - AC2S, ou seja:
 - **Estágio 1:** selecione amostra de m UPAs usando **AAS**.
 - **Estágio 2:** para cada UPA i da amostra de primeiro estágio, selecione n_i unidades secundárias das N_i unidades existentes usando **AAS**.

Para esse plano, a probabilidade de inclusão da unidade j da UPA i é dada por:

$$\pi_{ij} = P(i \in a, j \in s) = P(i \in a)P(j \in s | i \in a) = \frac{m}{M} \frac{n_i}{N_i}$$

Planos amostrais são mais simples quando as probabilidades de inclusão são constantes, isto é, $\pi_{ij} = n/N$, $\forall i$ e $\forall j$. Nestas condições, o plano amostral é dito *equiponderado* ou *autoponderado*.

Com o plano amostral AC2S, isto pode ser conseguido tomando $n_i \propto N_i$.

Amostragem conglomerada em dois estágios

AC2 com AAS nos 2 estágios

- O estimador não viciado do total sob o plano amostral **AC2S** é (@Cochran1977, Expressão 11.21):

$$\hat{Y}_{AC2S} = \frac{M}{m} \sum_{i \in a} \hat{T}_i$$

com $\hat{T}_i = \frac{N_i}{n_i} \sum_{j \in s_i} y_{ij}$ para toda UPA i .

- A variância do estimador não viciado do total sob o plano amostral **AC2S** é (@Cochran1977, Expressão 11.22):

$$\begin{aligned} Var_{AC2S} \left(\hat{T}_{AC2S} \right) &= M^2 \left(\frac{1}{m} - \frac{1}{M} \right) \frac{1}{M-1} \sum_{i \in C} \left(T_i - \overline{Y_C} \right)^2 \\ &\quad + \frac{M}{m} \sum_{i \in C} N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2 \end{aligned}$$

onde as parcelas do segundo membro representam as componentes da

Amostragem conglomerada em dois estágios

AC2 com AAS nos 2 estágios

Note que:

i) Se $m = M$ então, a 1ª componente da variância é nula, ou seja:

$$Var_{AC2S}(\hat{Y}_{AC2S}) = \sum_{i \in C} N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2 = V_{AES}(\hat{Y}_{AES})$$

e este plano amostral equivaleria ao de uma amostra estratificada, em que os conglomerados se tornaram estratos!

ii) Se $n_i = N_i$ ($\forall i = 1, 2, \dots, n$) então, a 2ª componente da variância é nula, ou seja:

$$V_{AC2S}(\hat{Y}_{AC2S}) = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) \frac{1}{M-1} \sum_{i \in C} (Y_i - \bar{Y}_C)^2 = V_{AC1S}(\hat{Y})$$

e este plano amostral equivaleria ao de uma amostra de conglomerados em um estágio simples.

De acordo com @Cochran1977, Expressão (11.24), um estimador não viciado da variância do estimador HT do total sob o plano amostral AC2S é dado por:

$$\hat{V}_{HT}(\hat{Y}_{AC2S}) = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) \frac{1}{M-1} \sum_{i \in C} (\hat{Y}_i - \bar{\hat{Y}}_C)^2$$

Amostragem conglomerada em dois estágios

AC2 com AAS nos 2 estágios

- A variância do estimador não viciado da média por unidade sob o plano amostral AC2S é dada por:

$$Var_{AC2S}(\bar{y}_{AC2S}) = \frac{1}{N^2} Var_{AC2S}(\hat{T}_{AC2S})$$

Um estimador não viciado da variância do estimador HT da média por unidade sob o plano amostral AC2S é dado por:

$$\hat{V}ar_{AC2S}(\bar{y}_{AC2S}) = \frac{1}{N^2} \hat{V}ar_{AC2S}(\hat{T}_{AC2S})$$


Para casa

- Continuar exercícios.
- Ler o capítulo 3 da apostila da Profa. Vanessa.
- Ler o capítulo 8 do livro 'Amostragem: Teoria e Prática Usando R'.
- Rever os slides.

Próxima aula

- Acompanhar o material no moodle.

Amostragem Sistemática

- Estimação.
- Laboratório de 

Muito obrigado!



Fonte: imagem do livro *Combined Survey Sampling Inference: Weighing of Basu's Elephants*.

Referências

- Amostragem: Teoria e Prática Usando o R
- **Elementos de Amostragem**, Bolfarine e Bussab.
- Cochran(1977)

Resumo da notação

Trabalho

Tópicos em Amostragem com Probabilidades Variáveis

e

Amostragens Complexas

- Escolher um tema dentre os tópicos que encerram o conjunto de disciplinas de Amostragem.
- Materiais disponíveis:
 - minicurso Sinape
 - minicurso Thomas
 - capítulos das nossas referências
- Apresentar:
 - problema
 - delineamento e estratégia
 - 10 slides? máximo
 - 15 min apresentacao + 5 min perguntas. presencial ou video