MAT02036 - Amostragem 2

Aula 05 - Amostragem Estratificada - Alocação Ótima

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 📀

Amostragem estratificada simples com alocação **proporcional** - AES_{pr}

- Com $n_h = nW_h$, a *média amostral simples* é **ENV** da média populacional na $AASs$

$$\overline{y}_{AES_{pr}} = \sum_{h=1}^{H} W_h \overline{y}_h = \sum_{h=1}^{H} W_h \frac{1}{n_h} \sum_{i \in s_h} y_i = \frac{1}{n} \sum_{h=1}^{H} \sum_{i \in s_h} y_i = \overline{y}$$

- A variância de $\overline{y}_{AES_{pr}}$, na **AASs** simplifica para

$$S^2_{\overline{y}_{AES_{pr}}} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h S^2_{h,y} \doteq \left(\frac{1}{n} - \frac{1}{N}\right) S^2_D,$$

e na **AASc**

$$Var_{AES_{pr}}\left(\overline{y}_{AES_{pr}}\right) = \frac{1}{n} \sum_{h=1}^{H} W_h Var_h = \frac{1}{n} Var_D$$

Aula passada 💾

- Amostragem estratificada **uniforme** - AES_{un}

$$n_h = \frac{n}{H} = k \quad \text{e} \quad f_h = \frac{k}{N_h}$$

assim

$$\bar{y}_{AES_{un}} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} W_h \frac{1}{k} \sum_{i \in s_h} y_i = \frac{1}{k} \sum_{h=1}^{H} W_h \sum_{i \in s_h} y_i.$$

Questões

- O estimador da média $\bar{y}_{AES_{un}}$ é **ENV** para $\overline{Y}$❓
  - Note que $\bar{y}_{AES_{un}} \neq \bar{y}$, a média amostral $\bar{y}$ é **ENV** para $\overline{Y}$ na **AESun**?
- Como fica a expressão de $V_{AES_{un}}\left(\bar{y}_{AES_{un}}\right)$ nesse caso❓

Aula passada

- Alocação da amostra no estrato h
  - **Proporcional**, $\quad n_h = nW_h$
  - **Uniforme**, $\quad n_h = n/H$

- Estimadores da média

Alocação	média
Proporcional	$\overline{y}_{AES_{pr}} = \overline{y}$
Uniforme	$\overline{y}_{AES_{un}} = \frac{1}{k}\sum_{h=1}^{H} W_h \sum_{i \in s_h} y_i$

- Suas variâncias

Alocação	Sob AASc	Sob AASs
Proporcional	$Var\left(\overline{y}_{AES_{pr}}\right) = \frac{1}{n}\sum_{h=1}^{H} W_h Var_h$	$Var(\overline{y}_{AES_{pr}}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{H} W_h S_h^2$
Uniforme	$Var(\overline{y}_{AES_{un}}) = \frac{1}{k}\sum_{h=1}^{H} W_h^2 Var_h$	$Var(\overline{y}_{AES_{un}}) = \frac{1}{k}\sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) S_h^2$

Alocação ótima

Alocação ótima

- A maioria das pesquisas convive com **restrições orçamentárias**.
- É sempre possível ganhar eficiência com uso de **AES** em comparação com uma **AAS** de igual tamanho.
- O caminho da **alocação proporcional não** é o caminho que **permite** obter **o maior ganho** de eficiência possível.
- É nesse contexto que foi desenvolvido o método de **alocação ótima** para amostras estratificadas simples (**AESot**).

Alocação ótima

Função custo

- Seja o **custo total** da pesquisa fixado em C unidades monetárias,
  - uma **função custo** descreve como C varia para diferentes n e alternativas de alocação da amostra nos estratos.
  - Considere uma **função custo linear** dada por

$$C = c_0 + \sum_{h=1}^{H} n_h c_h \quad \text{ou} \quad C' = C - c_0 = \sum_{h=1}^{H} n_h c_h$$

  - c_0 representa os custos fixos da pesquisa
  - $n_h c_h$ os custos que dependem efetivamente de cada estrato h.
- Na **AES** temos que $Var_{AES}\left(\overline{y}_{AES}\right)$ é mínima para C fixado ou C é mínimo para $V_{AES}\left(\overline{y}_{AES}\right)$ fixada.

Alocação ótima

Minimização da Variância

A **variância do estimador** da **média** populacional pode ser escrita como:

$$V_{AES}\left(\bar{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 S_{h,y}^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) = \sum_{h=1}^{H} W_h^2 S_{h,y}^2 / n_h - V_0,$$

onde $V_0 = \sum_{h=1}^{H} W_h^2 S_{h,y}^2 / N_h$.

- V_0 não depende de n_h, para minimizar $V_{AES}\left(\bar{y}_{AES}\right)$ basta encontrar valores de n_h que minimizem $\sum_{h=1}^{H} W_h^2 S_{h,y}^2 / n_h$.

- Técnicas de minimização de funções com restrições lineares, ex. o método dos multiplicadores de Lagrange. (❓)

- O resultado da minimização corresponde à *alocação ótima* dada por:

$$n_h = n \times \frac{W_h S_{h,y} / \sqrt{c}_h}{\sum_{k=1}^{H} W_k S_{k,y} / \sqrt{c}_k}, \ \forall\, h = 1, \ldots, H$$

Proposto por Neyman(1934), em seu artigo seminal que introduziu as bases da amostragem probabilística, definiu a amostragem estratificada e já indicava a

Alocação ótima

Uma via de **demonstrar o resultado** é utilizando a desigualdade de Cauchy-Scwartz.

1. **Minimizar** $Var_{AES}\left(\bar{y}_{AES}\right)$ para C' fixado ou C' para $V_{AES}\left(\bar{y}_{AES}\right)$ fixada é equivalente a minimizar o produto

$$Var_{AES}\left(\bar{y}_{AES}\right)C' = \left(\sum_{h=1}^{H}\frac{W_h^2 S_h^2}{n_h}\right)\left(\sum_{h=1}^{H} n_h c_h\right).$$

2. A **desigualdade** diz que $(\sum_h a_h^2)(\sum_h b_h^2) \geq (\sum_h a_h^2 b_h^2)$ sendo a igualde quando $\frac{b_h}{a_h} = k$ (constante).

3. Definindo $a_h = \frac{W_h S_h}{\sqrt{n_h}}$ e $b_h = \sqrt{n_h c_h}$ então $\frac{b_h}{a_h} = \frac{\sqrt{n_h c_h}}{\frac{W_h S_h}{\sqrt{n_h}}} = \frac{n_h\sqrt{c_h}}{W_h S_h} = k$, para todo $h = 1, \ldots, H$.

- Temos então que $Var_{AES}\left(\bar{y}_{AES}\right)C'$ mínimo quando

$$n_h = k\frac{W_h S_h}{\sqrt{c_h}}$$

Alocação ótima

4. Como $n = \sum_{h=1}^{H} n_h$ então

$$k = \frac{n}{\sum_{h=1}^{H} W_h S_h} / \sqrt{c_n}.$$

5. Substituindo em n_h no passo (3) temos

$$n_h = n \frac{\frac{W_h S_h}{c_h}}{\sum_{h=1}^{H} \frac{W_h S_h}{c_h}}.$$

- n_h na **AESot** é diretamente proporcional a $W_h S_h$ e inversamente proporcional a $\sqrt{c_n}$.

Sob **alocação ótima**, uma amostra maior será selecionada num estrato h sempre que:

a. O estrato tiver mais unidades, N_h grande.
b. A variabilidade no estrato for maior, $S_{h,y}$ grande.
c. O custo de amostragem no estrato for menor, c_h pequeno.

Alocação (ótima) de Neyman

- Quando $S_h = S^*$ e $c_h = c^*, \quad \forall \quad h = 1, 2, \ldots, H$, ambos constantes,

$$n_h = nN_h/N$$

  a **alocação ótima coincide** com a **alocação proporcional**.

- Entretanto, se apenas os custos de amostragem forem constantes ao longo dos estratos, $c_h = c^*, \forall\, h = 1, 2, \ldots, H$, então:

$$n_h = n \times \frac{N_h S_{h,y}}{\sum_{k=1}^{H} N_k S_{k,y}}$$

gerando a chamada *Alocação (Ótima) de Neyman*.

- Ex. pesquisas de estabelecimentos quando os **desvios padrões** $S_{h,y}$ **crescem com o tamanho das unidades**, maior variação em estabelecimentos maiores.

Alocação (ótima) de Neyman

- Se *amostragem estratificada simples com alocação de Neyman - AESne* é usada,
    - então o valor da **variância minimizada** para o estimador da média populacional é dado por:

$$Var_{AES_{ne}}\left(\overline{y}_{AES}\right) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h S_{h,y}\right)^2 - \frac{1}{N}\left(\sum_{h=1}^{H} W_h S_{h,y}^2\right)$$

    - O segundo termo à direita corresponde à *correção de população finita*.
    - A Expressão é obtida pela substituição de n_h na expressão da variância do estimador pela expressão do n_h da alocação de Neyman.
- E no plano **AESne** mas agora sob **AASc**❓

Alocação (ótima) de Neyman

- sob **AASc** dentro dos estratos

$$
\begin{aligned}
Var_{AES_{ne}}\left(\bar{y}_{AES}\right) &= Var_{AES_{ne}}\left(\sum_{h=1}^{H} W_h \bar{y}_h\right) \\
\text{(❓)} &= \sum_{h=1}^{H} W_h^2 Var_{AES_{ne}}\left(\bar{y}_h\right) \\
\text{(def. AASc dentro)} &= \sum_{h=1}^{H} W_h^2 \frac{Var_h}{n_h} \\
(n_h \text{ na } AES_{ne}) &= \sum_{h=1}^{H} W_h^2 \frac{Var_h}{n \frac{\sqrt{Var_h} N_h}{\sum_{k=1}^{H} \sqrt{Var_k} N_k}} = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{Var_h}{n} \frac{\sum_{k=1}^{H} \sqrt{Var_k} N_k}{\sqrt{Var_h} N_h} \\
&= \frac{1}{n} \sum_{h=1}^{H} W_h \sqrt{Var_h} \sum_{k=1}^{H} \frac{\sqrt{Var_k} N_k}{N} = \frac{1}{n} \sum_{h=1}^{H} W_h \sqrt{Var_h} \sum_{k=1}^{H} W_k \sqrt{Var_k} \\
\text{(❓)} &= \frac{1}{n}\left(\sum_{h=1}^{H} W_h \sqrt{Var_h}\right)^2 = \frac{\overline{DP}}{n}
\end{aligned}
$$

para $DP_h = \sqrt{Var_h}$ e $\overline{DP} = \sum_{h=1}^{H} W_h DP_h$.

Exemplo 🏃

Exercício 4.1 do livro "Elementos de Amostragem" (Bolfarine e Bussab)

Uma população está dividida em 5 estratos. Os tamanhos dos estratos N_h, médias $\overline{Y}$ e variâncias S_h^2 são dados na tabela abaixo.

h	Nh	$\overline{Y}$	S_h^2
1	117	7,3	1,31
2	98	6,9	2,03
3	74	11,2	1,13
4	41	9,1	1,96
5	45	9,6	1,74

a. Calcule os parâmetros globais $\overline{Y}$ e Var_y.
b. Para uma amostra de tamanho $n = 80$, determine as alocações proporcional e (ótima) de Neyman.
c. Compare as variâncias dos estimadores obtidos sob **AASc** e **AESne**.
d. Faça o mesmo para a **AASc** e a **AESpr**.

Para casa 🏠

- Continuar o Exemplo.
- Mostrar $V_{AES}\left(\overline{y}_{AES_{ne}}\right) \leq V_{AES_{pr}}\left(\overline{y}_{AES}\right) \leq V_{AAS}\left(\overline{y}\right)$ sob **AASc** dentro dos estratos.
- Fazer exercício 11.5 do livro 'Amostragem: Teoria e Prática Usando R' https://amostragemcomr.github.io/livro/estrat.html#exerc11
- Fazer exercício 1 da lista 1.
- Rever os slides.

Próxima aula 📊

- Amostragem Estratificada
  - Mais sobre Comparação de alternativas de alocação da amostra, efeito do Plano Amostral
- Laboratório de R

Muito obrigado!

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

Resumo da notação

Alocação ótima

- **Função custo linear**: $C = c_0 + \sum_{h=1}^{H} n_h c_h$ ou $C' = C - c_0 = \sum_{h=1}^{H} n_h c_h$.

Alocação	sob AASc dentro	sob AASs dentro
Ótima	$n_h = n \times \frac{W_h \sqrt{Var_{h,y}}/\sqrt{c_h}}{\sum_{k=1}^{H} W_k \sqrt{Var_{k,y}}/\sqrt{c_k}}$	$n_h = n \times \frac{W_h S_{h,y}/\sqrt{c_h}}{\sum_{k=1}^{H} W_k S_{k,y}/\sqrt{c_k}}$
de Neyman	$n_h = n \times \frac{N_h \sqrt{Var_{h,y}}}{\sum_{k=1}^{H} N_k \sqrt{Var_{k,y}}}$	$n_h = n \times \frac{N_h S_{h,y}}{\sum_{k=1}^{H} N_k S_{k,y}}$

- Variâncias na **AESne**

Plano dentro	Variância $\bar{y}_{AES}$ na AESne
AASc	$Var_{AES_{ne}}\left(\bar{y}_{AES}\right) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h \sqrt{Var_h}\right)^2 = \frac{\overline{DP}^2}{n}$
AASs	$Var_{AES_{ne}}\left(\bar{y}_{AES}\right) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h S_{h,y}\right)^2 - \frac{1}{N}\left(\sum_{h=1}^{H} W_h S_{h,y}^2\right)$

em que $DP_h = \sqrt{Var_h}$ e $\overline{DP} = \sum_{h=1}^{H} W_h DP_h$.

Referências

Slides baseados no Capítulo 11 do livro

- Amostragem: Teoria e Prática Usando o R

Citações do Capítulo

- Neyman(1934)
- Cochran(1977)