MAT02036 - Amostragem 2

Aula 06 - Amostragem Estratificada - Mais sobre Alocações e comparações

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

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 $\overline{DP} = \sum_{h=1}^{H} W_h DP_h$.

Aula passada 💾

Exercício 4.1 (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**.

Exercício 4.1 (Bolfarine e Bussab) 💪

Dados do problema:

```
H <- 5                                         # no. de estratos
h <- 1:H                                       # indice dos estratos
Nh <- c( 117, 98, 74, 41, 45)                  # tamanho dos estratos
Ybarrah <- c( 7.3, 6.9, 11.2, 9.1, 9.6)        # media pop. dos estratos
S2h <- c( 1.31, 2.03, 1.13, 1.96, 1.74)        # variancia do estrato
N <- sum(Nh)                                   # tamanho da populacao
n <- 80                                        #tamanho de amostra
```

a. No R temos (ver expressões nos próximos slides)

```
## a.
Ybarra <- sum( Nh * Ybarrah) / N         # media pop global
Ybarra
```

```
## [1] 8.437867
```

```
Vary_aux1 <- sum((Nh - 1) * S2h) / N                # primeiro term
Vary_aux2 <- ( sum( Nh * Ybarrah^2) / N) - Ybarra^2  # segundo termo
Vary <- Vary_aux1 - Vary_aux2                        # variancia pop
Vary
```

Exercício 4.1 (Bolfarine e Bussab)🏃‍♀️

Seguindo as **expressões** vistas em aula a média global é dada por

$$\overline{Y} = \frac{1}{N}\sum_{i\in U} y_i = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in U_h} y_i = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{\sum_{i\in U_h} y_i}{N_h} = \sum_{h=1}^{H}\frac{N_h}{N}\frac{\sum_{i\in U_h} y_i}{N_h} = \sum_{h=1}^{H} W_h \overline{Y}_h$$
$$= \frac{117\times 7.3 + 98\times 6.9 + 74\times 11.2 + 41\times 9.1 + 45\times 9.6}{375} = 8.438$$

e para a variância sabemos que

$$Var_y = \frac{1}{N}\sum_{i\in U}\left(y_i - \overline{Y}\right)^2 = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\left(y_i - \overline{Y}\right)^2 = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\left(y_i - \overline{Y}_h + \overline{Y}_h - \overline{Y}\right)^2$$
$$= \frac{1}{N}\sum_{h=1}^{H} N_h Var_{h,y} + \frac{1}{N}\sum_{h=1}^{H} N_h\left(\overline{Y}_h - \overline{Y}\right)^2 .$$

No primeiro termo, note que $Var_{h,y} = \frac{N_h - 1}{N_h} S_h^2$, então

$$\frac{1}{N}\sum_{h=1}^{H} N_h Var_{h,y} = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{N_h - 1}{N_h} S_h^2 = \frac{1}{N}\sum_{h=1}^{H}(N_h - 1) S_h^2$$
$$= \frac{116\times 1.31 + 97\times 2.03 + 73\times 1.13 + 40\times 1.96 + 44\times 1.74}{N} = 1.5635 .$$

Exercício 4.1 (Bolfarine e Bussab) 🏋

No segundo termo, podemos mostar (❓) que

$$\frac{1}{N}\sum_{h=1}^{H} N_h\left(\overline{Y}_h - \overline{Y}\right)^2 = \ldots = \left(\frac{1}{N}\sum_{h=1}^{H} N_h \overline{Y}_h^2\right) - \overline{Y}^2$$
$$= \frac{117 \times 7.3^2 + 98 \times 6.9^2 + 74 \times 11.2^2 + 41 \times 9.1^2 + 45 \times 9.6^2}{375} - 8.438^2$$
$$= 73.9351 - 71.1976 = 2.7376$$

Então

$$Var_y = 1.5635 + 2.7376 = 4.3011$$

b. No R temos (ver expressões nos próximos slides)

```
## b.
nhpr <- n * Nh / N                                      # vetor de nh's
nhpr
```

```
## [1] 24.960000 20.906667 15.786667  8.746667  9.600000
```

```
nhne <- n * (Nh * sqrt(S2h)) / (sum(Nh * sqrt(S2h)))   # vetor de nh's
nhne
```

Exercício 4.1 (Bolfarine e Bussab)🏃

No plano **AESpr** temos $n_h = n \times W_h$, em que $W_h = N_h/N$, assim

$$n_1 = 80 \times \frac{117}{375}, \ldots, n_5 = 80 \times \frac{45}{375}$$

No plano **AESne** temos $n_h = n \times \frac{N_h S_{h,y}}{\sum_{k=1}^{H} N_k S_{k,y}}$, calculando primeiro o denominador

$$\begin{aligned}
\sum_{k=1}^{H} N_k S_{k,y} &= \sum_{k=1}^{H} N_k \sqrt{S_{k,y}^2} \\
&= 117 \times \sqrt{1.31} + 98 \times \sqrt{2.03} + 74 \times \sqrt{1.13} + 41 \times \sqrt{1.96} + 45 \times \sqrt{1.74} \\
&= 473
\end{aligned}$$

e assim

$$n_1 = 80 \times \frac{117 \times \sqrt{1.31}}{473}, \ldots, n_5 = 80 \times \frac{45 \times \sqrt{1.74}}{473}$$

c. continuar...
d. continuar...

Comparação de alternativas de alocação da amostra

Comparação de alternativas de alocação da amostra

- Particionando a **soma de quadrados total** em parcelas devidas à **variação dentro e entre** estratos (e ignorando termos de ordem $1/N_h$),
  - sob *alocação de Neyman*, asssumindo

AASs dentro dos estratos	ou AASc dentro dos estratos
$n_h \propto N_h S_{h,y}$	$n_h \propto N_h Var_{h,y}$

pode-se mostrar que (Cochran, 1977; página 99):

$$V_{AESne}\left(\overline{y}_{AES}\right) \leq V_{AESpr}\left(\overline{y}_{AES}\right) \leq V_{AAS}\left(\overline{y}\right)$$

- **AES** com alocação de **Neyman** é **mais eficiente** que **AES** com alocação **proporcional**.

- Ambas superam **AAS** como plano amostral para um mesmo tamanho especificado de amostra.

Comparação de alternativas de alocação da amostra

Para o estimador da **média**, assumindo **AASc** dentro dos estratos, temos que

1. Usando a partição da variâcia global de Y e **ignorando os estratos**,

$$Var_{AAS_c}(\overline{y}) = \frac{Var_y}{n} = \frac{Var_D}{n} + \frac{Var_E}{n}.$$

2. Na **AESpr** A variâcia do estimador da média, $Var_{AES_{pr}} = \frac{Var_D}{n}$ então,

$$Var_{AAS_c}(\overline{y}) = Var_{AES_{pr}} + \frac{Var_E}{n}, \text{ sendo que } \frac{Var_E}{n} \geq 0.$$

3. Na **AESne** temos
$Var_{AES_{ne}}\left(\overline{y}_{AES}\right) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h \sqrt{Var_h}\right)^2 = \frac{1}{n}\left(\sum_{h=1}^{H} W_h DP_h\right)^2 = \frac{\overline{DP}^2}{n}.$

Escrevendo $Var_{AES_{pr}} = \sum_{h=1}^{H} W_h Var_h = \sum_{h=1}^{H} W_h (DP_h)^2$ temos

$$\begin{aligned} Var_{AES_{pr}}\left(\overline{y}_{AES}\right) - Var_{AES_{ne}}\left(\overline{y}_{AES}\right) &= \frac{1}{n}\left\{\sum_{h=1}^{H} W_h (DP_h)^2 - \left(\sum_{h=1}^{H} W_h DP_h\right)^2\right\} \\ &= \frac{1}{n}\sum_{h=1}^{H} W_h \left(DP_h - \overline{DP}\right)^2 = \frac{Var_{DP}}{n} \end{aligned}$$

Comparação de alternativas de alocação da amostra

4. O termo $Var_{AES_{pr}}\left(\overline{y}_{AES}\right) - Var_{AES_{ne}}\left(\overline{y}_{AES}\right) = \frac{Var_{DP}}{n}$ representa a variabilidade entre os desvios padrões entre os estratos.

Então, fazendo

$$Var_{AES_{pr}}\left(\overline{y}_{AES}\right) = Var_{AES_{ne}}\left(\overline{y}_{AES}\right) + \frac{Var_{DP}}{n},$$

mostramos

$$Var_{AAS_c}(\overline{y}) = Var_{AES_{pr}} + \frac{Var_E}{n} = Var_{AES_{ne}}\left(\overline{y}_{AES}\right) + \frac{Var_{DP}}{n} + \frac{Var_E}{n}.$$

- Sempre que os estratos tem **médias distintas**, $\frac{Var_E}{n}$ grande, **AESpr** e **AESne** serão **vantajosas**.
- Se os devios padrões dos estratos também diferirem muito, Var_{DP} grande, recomenda-se a **AESne**.

Alguns problemas com alocação ótima

1.. Em geral, os valores de $S_{h,y}$, $h = 1, \ldots, H$, são desconhecidos.
a. Usar informações de uma variável auxiliar x, usando $S_{h,x}$.
b. Predizer y_i usando informações auxiliares x_i, e então estimar $S_{h,y}$ a partir dos valores preditos.
c. Usar o total ou a amplitude da variável auxiliar x no estrato h como *proxy* para $S_{h,y}$.
d. Selecionar pequena amostra piloto (preliminar) e usar dados desta amostra para estimar $S_{y,h}$.

2. Pode haver muitas variáveis de pesquisa y.
a. Usar a média das alocações alternativas em cada estrato.
b. Escolher uma ou duas variáveis principais, média das alocações.
c. Construir um 'índice' das variáveis e usar para definir a alocação.
d. Usar alocação proporcional.

3. Se $n_h > N_h$ para algum estrato.

- Fazer $n_h = N_h$, **estrato certo** ou **estrato censitário**, se $n_h > N_h$.
- Em seguida, refazer a alocação ótima nos demais estratos e ajustar o tamanho da amostra.
- @Brito2015 oferecem uma solução exata utilizando uma formulação de *Programação Inteira Binária*. o pacote *stratbr* para o R está disponível - ver @Brito2019 para detalhes.

Alguns problemas com alocação ótima

4. Se $n_h < 2$ para algum estrato.

- Se estimar variâncias importa, forçar $n_h \geq 2$ para todo h.
- Na prática, se usa $n_h \geq 5$ devido à possibilidade de **não resposta**. Estimar sem viés o total ou média necessita $n_h \geq 2$ para todos h.
- Se algum $n_h = 1$, utilizar métodos aproximados para estimação de variâncias, tais como agregação de estratos ou similares (ver @Cochran1977, Seção 5A.12).

5. Ganhos de eficiência podem ser modestos, particularmente na estimação de proporções. (@Cochran1977, página 99)

$$V_{AESN}\left(\bar{y}_{AES}\right) \leq V_{AESP}\left(\bar{y}_{AES}\right) \leq V_{AAS}\left(\bar{y}\right)$$

- Ganhos de precisão dependem da relação entre a(s) variável(is) de estratificação e as variáveis de pesquisa.
- Em geral, os **ganhos** são **pequenos** para amostras de pessoas e variáveis ligadas a atitudes, opiniões, comportamentos, etc.
- Para pesquisas amostrais de estabelecimentos ou instituições, os **ganhos** podem ser **maiores**.

se os ganhos de precisão alcançados com a estratificação não são grandes, o responsável pelo planejamento da pesquisa precisa avaliar se a estratificação

Efeito do Plano Amostral (EPA)

- Também chamado Efeito de Delineamento, em inglês $deff$ (*Design Effect*)
  - Seja *plano* um plano amostral

$$EPA_{plano} = deff_{plano} = \frac{Var(\bar{y}_{plano})}{Var(\bar{y}_{AAS_c})}.$$

- se $deff_{plano} < 1$ então o **plano** é mais eficiente que a **AASc**.

Exemplo: 🏋️

Sabemos mostrar $EPA_{AES_{pr}} = deff_{AES_{pr}}$ e $EPA_{AES_{pr}} = deff_{AES_{pr}}$ assumindo **AASc** dentro dos estratos❓

Efeito do Plano Amostral (EPA)

Já havíamos falado sobre efeito de planejamento

Exemplo 6 da Apostila da Profa Vanessa: 🏃‍♀️

Seja os dados do exemplo 1, da população de 8 domicílios.

(trabalhamos nos dados do exemplo 1 nos nossos slides **Aula 03** e **Aula 04**)

Para casa 🏠

- Continuar os Exemplos.
- Mostrar tamanho de amostra n para AASc dentro dos estratos.
- Fazer exercícios 11.7 e 11.10 do livro 'Amostragem: Teoria e Prática Usando R' https://amostragemcomr.github.io/livro/estrat.html#exerc11
- Fazer exercícios :: da lista 1.
- Rever os slides.

Próxima aula 📊

- Amostragem Estratificada
    - Estimação de proporções
    - Exercícios e Intervalos de confiança

Muito obrigado!

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

Resumo da notação

Para o estimador da **média**, assumindo **AASc** dentro dos estratos

$$Var_{AAS_c}(\overline{y}) = Var_{AES_{pr}} + \frac{Var_E}{n} = Var_{AES_{ne}}\left(\overline{y}_{AES}\right) + \frac{Var_{DP}}{n} + \frac{Var_E}{n}.$$

- Efeito do Plano Amostral/Delineamento (*Design Effect*)
  - Seja *plano* um plano amostral

$$EPA_{plano} = deff_{plano} = \frac{Var(\overline{y}_{plano})}{Var(\overline{y}_{AAS_c})}.$$

Referências

Slides baseados no Capítulo 11 do livro

- Amostragem: Teoria e Prática Usando o R

Citações do Capítulo

- Neyman(1934)
- Cochran(1977)