MAT02036 - Amostragem 2

Aula 04 - Amostragem Estratificada - Propriedades e Alocação de amostras

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

Parâmetros no estrato h

- Total - $T_h = \sum_{i \in U_h} y_i$
- Média - $\overline{Y_h} = T_h / N_h = \frac{1}{N_h} \sum_{i \in U_h} y_i$
- Variância - $S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_i - \overline{Y_h})^2 \quad \text{ou} \quad Var_h = Var_{h,y} = \frac{N_h - 1}{N_h} S_h^2$

Parâmetros globais

- Total - $T = \sum_{h=1}^{H} T_h = \sum_{h=1}^{H} N_h \overline{Y_h}$
- Média - $\overline{Y} = T/N = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{Y_h} = \sum_{h=1}^{H} W_h \overline{Y}_h, W_h = N_h / N$
- Variância - $S^2 = \frac{1}{N-1} \sum_{h=1}^{H} \sum_{i \in U_h} \left(y_i - \overline{Y} \right)^2 \quad \text{ou} \quad Var_y = \frac{1}{N} \sum_{h=1}^{H} (N_h - 1) S_h^2$

Decomposição da variância populacional

$$S_y^2 = \frac{1}{N-1} \sum_{h=1}^{H} (N_h - 1) S_h^2 + \frac{1}{N-1} \sum_{h=1}^{H} N_h \left(\overline{Y_h} - \overline{Y} \right)^2 = S_D^2 + S_E^2$$

ou

$$Var_y = \sum_{h=1}^{H} W_h Var_h + \sum_{h=1}^{H} W_h \left(\overline{Y_h} - \overline{Y} \right)^2 = Var_D + Var_E.$$

Aula passada

Estimadores no estrato h

Estimadores AASc	Estimadores AASs
$\widehat{T}_h = \frac{N_h}{n_h} \sum_{i \in s_h} y_i = N_h \overline{y}_h$	$\widehat{T}_h = \frac{N_h}{n_h} \sum_{i \in s_h} y_i = N_h \overline{y}_h$
$\overline{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i$	$\overline{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i$
$Var_{AESc}\left(\widehat{T}_h\right) = N_h^2 Var_{AESc}\left(\overline{y}_h\right)$	$Var_{AES}\left(\widehat{T}_h\right) = N_h^2 Var_{AES}\left(\overline{y}_h\right)$
$Var_{AESc}\left(\overline{y}_h\right) = \frac{1}{n_h} Var_h$	$Var_{AES}\left(\overline{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$
$\widehat{Var}_{AESc}\left(\widehat{T}_h\right) = N_h^2 \widehat{Var}_{AESc}\left(\overline{y}_h\right)$	$\widehat{Var}_{AES}\left(\widehat{T}_h\right) = N_h^2 \widehat{Var}_{AES}\left(\overline{y}_h\right)$
$\widehat{Var}_{AESc}\left(\overline{y}_h\right) = \frac{1}{n_h} \widehat{S}_h^2$	$\widehat{Var}_{AES}\left(\overline{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$

em que $Var_h = Var_h(y) = \frac{N_h - 1}{N_h} S_h^2$ e $\widehat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \left(y_i - \overline{y}_h\right)^2$.

Aula passada

Estimadores globais

- Do total T: $\widehat{T}_{AES} = \sum_{h=1}^{H} \widehat{T}_h = \sum_{h=1}^{H} N_h \overline{y}_h.$
- Da média $\overline{Y}$: $\overline{y}_{AES} = \sum_{h=1}^{H} W_h \overline{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h.$

Variância do estimador e seu **estimador**

Sob AASc	Sob AASs
$Var_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \frac{Var_h}{n_h}$	$Var_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$
$Var_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \frac{Var_h}{n_h}$	$Var_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$
$\widehat{Var}_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \frac{\widehat{S}_h^2}{n_h}$	$\widehat{Var}_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$
$\widehat{Var}_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \frac{\widehat{S}_h^2}{n_h}$	$\widehat{Var}_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$

Ainda aula passada 🏠

Tarefa para casa:

Seja $\widehat{T}_h$ um estimador não viesado (**ENV**) para T_h, o total do estrato h, segundo um plano amostral A, ou seja, $E_A(\widehat{T}_h) = T_h$, $\forall h = 1, \ldots, H$. De acordo com o plano amostral estratificado simples **AES** responda:

1. Mostrar $E_{AES}\left(\widehat{T}_{AES}\right) = T$, em que $\widehat{T}_{AES} = \sum_{h=1}^{H} N_h^2 \widehat{T}_h$;

2. Encontre $Var_{AES}\left(\widehat{T}_{AES}\right)$, tanto para **AASs** quanto para **AASc**.

Alguém tentou❓❓❓

Ainda aula passada 🏃‍♀️

Exercício: Para os dados do exemplo da aula passada (ver abaixo);

1. selecione uma amostra asumindo o plano $AASc$ em cada estrato;

2. construa um intervalo de 99% de confiança para a renda média global dos domicílios;

3. construa um intervalo de 99% de confiança para a renda média dos domicílios por estrato.

Iremos explorar os ICs nas próximas aulas ❗

Alocação da amostra pelos estratos

Exemplo aula passada 🏃 Alocação UNIFORME

Considere uma pesquisa feita em uma população de 8 domicílios, onde são conhecidas as variáveis renda familiar (Y) e local do domicílio (L), com os códigos A para região alta e B para baixa (B).

```
N <- 8
domicilio <- 1:N
y <- c( 13, 17, 6, 5, 10, 12, 19, 6)
l <- c( "B", "A", "B", "B", "B", "A", "A", "B")
```

domicilio	1	2	3	4	5	6	7	8
y	13	17	6	5	10	12	19	6
l	B	A	B	B	B	A	A	B

1. Calcule $Var(\overline{y})$ sob AASc, para estimar $\overline{Y} = \frac{T}{N} = \frac{\sum_i \in U y_i}{N}$, com $n = 4$.

2. Calcule $Var(\overline{y}_{AES})$, para $\overline{y}_{AES} = \frac{N_A \, \overline{y}_A + N_B \, \overline{y}_B}{N}$ (usando L como variável estratificadora), com amostra $n_h = 2$ por estrato.

3. Compare as variâncias dos dois planos (Efeito do Plano Amostral EPA).

Exemplo aula passada Alocação UNIFORME

Resolução no R:
$N = 8$ domicílios, $N_A = 3$, $N_B = 8$;
L: localidade domicílio, $H = 2$;
Y: renda familiar.

1. Sob AASc $Var(\bar{y}_{AASc}) = Var_y/n$,

```r
n <- 4                              # tamanho da amostra
Ybarra <- mean(y)                   # media populacional
vary <- sum((y-Ybarra)^2) / N       # variancia de Y_U
varybarra <- vary / n               # variancia da media amostral
```

2. $Var(\bar{y}_{AES}) = \sum_{h=1}^{H} W_h^2 Var(\bar{y}_h)$, então precisamos $Var(\bar{y}_A)$ e $Var(\bar{y}_B)$,

```r
yA <- y[l=="A"]
N_A <- length(yA)
nA <- 2
YbarraA <- mean(yA)
varyA <- sum((yA-YbarraA)^2)/N_A
varybarraA <- varyA / nA
```

```r
yB <- y[l=="B"]
N_B <- length(yB)
nB <- 2
YbarraB <- mean(yB)
varyB <- sum((yB-YbarraB)^2)/N_B
varybarraB <- varyB / nB
```

Exemplo aula passada 🏃 Alocação UNIFORME

Resolução no R:

Combinando as estimativas dos estratos, $Var(\overline{y}_{AES_{un}}) = \sum_{h=1}^{2} \left(\frac{N_h}{N}\right)^2 \overline{y}_h =$

```
(varybarraAES <- (N_A/N)^2 * varybarraA + (N_B/N)^2 * varybarraB)
```

```
## [1] 2.40625
```

3️⃣ Então $EPA = \frac{Var(\overline{y}_{AES})}{Var(\overline{y}_{AASc})} =$

```
(EPA <- varybarraAES / varybarra)
```

```
## [1] 0.4010417
```

Portanto, a **variância** do estimador foi **reduzida em mais da metade**, com o **mesmo tamanho amostral**❗❗❗

(cont.) Exemplo aula passada 🏃‍♀️

"Alocação PROPORCIONAL"

Com os mesmos dados do exemplo da aula passada, definimos outra estratégia de AES tal que os tamanhos amostrais são diferentes. Assuma agora que $n_A = 1$ e $n_B = 3$, com **AASc** dentro de cada estrato.

1. Encontre $Var_{AES}(\overline{y}_{AES})$.

2. Compare com os resultados de $Var(\overline{y})$ baseados nas estratégias anteriores, **AASc** com n=4 e **AES** com $n_A = n_B = 2$.

3. Como ficaria $Var_{AES}(\widehat{T})$ para o plano $AASs$ dentro de cada estrato?

(cont.) Exemplo aula passada 💪

Alocação "PROPORCIONAL"

Resolução no R:

1. Vimos que $Var(\overline{y}_{AES}) = \sum_{h=1}^{H} W_h^2 Var(\overline{y}_h)$, então precisamos $Var(\overline{y}_A)$ e $Var(\overline{y}_B)$,

```
yA <- y[l=="A"]
N_A <- length(yA)
nA <- 1
YbarraA <- mean(yA)
varyA <- sum((yA-YbarraA)^2)/N_A
varybarraA <- varyA / nA
```

```
yB <- y[l=="B"]
N_B <- length(yB)
nB <- 3
YbarraB <- mean(yB)
varyB <- sum((yB-YbarraB)^2)/N_B
varybarraB <- varyB / nB
```

Combinando as estimativas em cada estrato, $Var(\overline{y}_{AES_{pr}}) = \sum_{h=1}^{2} \left(\frac{N_h}{N}\right)^2 \overline{y}_h =$

```
(varybarraAESpr <- (N_A/N)^2 * varybarraA + (N_B/N)^2 * varybarraB)
```

```
## [1] 2.416667
```

Exemplo aula passada 🏋

Alocação "PROPORCIONAL"

Resolução no R:

2. Então $EPA = \frac{Var(\bar{y}_{AES_{pr}})}{Var(\bar{y}_{AES_{un}})} =$

```
(EPA <- varybarraAESpr / varybarraAES)
```

```
## [1] 1.004329
```

Nesse caso parece não haver grande diferença na redução da variância comparando os planos AES_{un} e AES_{pr}, mas ambos mostram que a variância do estimador foi **reduzida em mais da metade**, em relação a uo plano $AASc$, com o **mesmo tamanho amostral**‼️❗

Note que não usamos exatamente um plano proporcional, $\frac{n_A}{n} = \frac{1}{4} \neq \frac{3}{8} = \frac{N_A}{N}$.

3. *...continuar...*

Alocação da amostra nos estratos

- Uma decisão importante é a forma pela qual o tamanho total da amostra será alocado ou distribuído nos estratos.
- Em *estratos naturais* pode ser de **interesse calcular tamanhos de amostra** para que a estimação de parâmetros dos estratos tenham precisão controlada.
    - O tamanho total da amostra é obtido somando os tamanhos de amostra calculados para os estratos,
    - a alocação nos estratos vem antes da obtenção do tamanho total da amostra.
- Se **não há interesse específico na estimação** de parâmetros dos **estratos** e um tamanho total de amostra foi calculado, é necessário distribuir esse tamanho entre os estratos definidos na população.
- Há duas maneiras principais de alocação da amostra, que pode ser feita de forma *proporcional* ou *desproporcional* aos tamanhos N_h dos estratos.

Alocação Proporcional

Alocação Proporcional

Amostragem estratificada simples com alocação proporcional- AESpr$

- A **fração amostral** em cada **estrato**, $f_h = n_h/n$, é **constante** e igual à fração amostral da amostra inteira, $f = N_h/N$.
  - Estratos maiores ficam com amostras maiores.
- A **Alocação Proporcional** implica tentar implementar

$$\frac{n_h}{n} = \frac{N_h}{N} \Rightarrow n_h = n\frac{N_h}{N} = nW_h, \;\; \forall h = 1, 2, \ldots, H.$$

- Os tamanhos de amostra calculados provavelemente não serão números inteiros.
  - Na prática **arredondar** para cima os tamanhos de amostra calculados implica um pequeno aumento em n.
- A amostra sorteada será **auto-ponderada** e o procedimento de estimação poderá ser simplificado.
  - Todas as outras formas de alocação vão resultar em uma alocação *desproporcional* da amostra nos estratos.

Alocação Proporcional

- Com $n_h = nW_h$, temos que $\bar{y}_{AES_{pr}} = \bar{y}$ é **ENV** da média populacional,

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

- Na **AESpr** Sob **AASs** a **variância** de $\bar{y}_{AES_{pr}}$ simplifica para

$$Var_{AES_{pr}}(\bar{y}_{AES_{pr}}) \doteq \left(\frac{1}{n} - \frac{1}{N} \right) S_D^2$$

- A expressão aproximada tem a mesma forma que no caso da **AASs**, com S_y^2 substituído por S_D^2.
    - Como $\left(S_D^2 < S_y^2\right)$, a AES_{pr} geralmente reduz a variância do estimador se comparada com **AASs** de igual tamanho.

- Ou, na **AASc** temos

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

Alocação Proporcional

Podemos verificar que $\frac{W_h^2}{n_h} = \frac{W_h^2}{nW_h} = \frac{W_h}{n}$ e $\frac{W_h^2}{N_h} = \frac{W_h^2}{NW_h} = \frac{W_h}{N}$

- Então na **AASs** temos

$$
\begin{aligned}
Var_{AES_{pr}}(\bar{y}_{AES_{pr}}) &= Var_{AES_{pr}}\left[\sum_{h=1}^{H} W_h \bar{y}_h\right] \\
\text{❓} &= \sum_{h=1}^{H} W_h^2 Var_{AES_{pr}}\left(\bar{y}_h\right) \\
(\text{def. } \bar{y}_h \text{ na AASs}) &= \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 \\
(\text{igualdades acima}) &= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h S_{h,y}^2 \\
&\doteq \left(\frac{1}{n} - \frac{1}{N}\right) S_D^2,
\end{aligned}
$$

onde a aproximação no último termo decorre de usar pesos N_h/N em lugar dos pesos $(N_h - 1)/(N - 1)$ como na definição de S_D^2.

- Na **AASc**❓❓❓

Alocação Uniforme

Alocação Uniforme

Amostragem estratificada uniforme- AESun

- O **tamanho amostral** em cada **estrato** é **constante**.
- Procedimento indicado quando se pretende **apresentar estimativas separadas** para cada estrato.
- Na **Alocação Uniforme** temos

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

assim

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

Questões

- O estimador da média $\overline{y}_{AES_{un}}$ é **ENV** para $\overline{Y}$ ❓
    - Note que $\overline{y}_{AES_{un}} \neq \overline{y}$, a média amostral $\overline{y}$ é **ENV** para $\overline{Y}$ na **AESun**?

Alocação Uniforme

Com $n_h = \frac{n}{H} = k$ e $f_h = \frac{n_h}{N_h} = \frac{k}{N_h}$ temos $N_h = \frac{kN_h}{n_h}$

- sob **AASs** temos

$$
\begin{aligned}
Var_{AES_{un}}(\bar{y}_{AES_{un}}) &= Var_{AES_{pr}}\left[\sum_{h=1}^{H} W_h \bar{y}_h\right] \\
❓ &= \sum_{h=1}^{H} W_h^2 Var_{AES_{un}}\left(\bar{y}_h\right) \\
(\text{def. } \bar{y}_h \text{ na AASs}) &= \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 \\
(\text{igualdades acima}) &= \sum_{h=1}^{H} W_h^2 \left(\frac{1}{k} - \frac{1}{\frac{kN_h}{n_h}}\right) S_h^2 \\
&= \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{k},
\end{aligned}
$$

- Na **AASc** $Var_{AES_{un}}(\bar{y}_{AES_{un}}) = \sum_{h=1}^{H} W_h^2 \frac{Var_h}{k}$
  - Sabemos mostrar❓❓❓

Para casa 🏠

- Continuar o Exemplo.
- Mostrar $E_{AES}\left(\widehat{T}_{AES}\right) = T$.
- Fazer exercícios 11.1 e 11.2 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.
- Ler a partir da seção 11.3 do livro 'Amostragem: Teoria e Prática Usando R'.

Próxima aula 📊

- Amostragem Estratificada
    - continuação alocação de amostras

Muito obrigado!

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

Resumo da notação

- Tamanho 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	$Var(\overline{y}_{AES_{un}}) = \frac{1}{k}\sum_{h=1}^{H} W_h^2 Var_h$	$Var(\overline{y}_{AES_{un}}) = \frac{1}{k}\sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) S_h^2$

Referências

Slides baseados no Capítulo 11 do livro

- Amostragem: Teoria e Prática Usando o R

Citações do Capítulo

- Horsfield(2017)
- IBGE(2000)