MAT02036 - Amostragem 2

Aula 08 - Amostragem Estratificada - Exercícios e Lab R

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

- **Tamanho da amostra** na **AES** dado um tipo de **alocação**, w_h, e **fixando a variância** máxima que se deseja para a estimativa do parâmetro, V

Parâmetro	Sob AASc dentro dos estratos	Sob AASs dentro dos estratos
Média	$n \geq \frac{\sum_{h=1}^{H} W_h^2 \frac{Var_{h,y}}{w_h}}{V}$	$n \geq \frac{\sum_{h=1}^{H} \frac{W_h^2 S_{h,y}^2}{w_h}}{V + \frac{1}{N}\sum_{h=1}^{H} W_h S_{h,y}^2}$
Total	$n \geq \frac{\sum_{h=1}^{H} N_h^2 \frac{Var_{h,y}}{w_h}}{V}$	$n \geq \frac{\sum_{h=1}^{H} \frac{N_h^2 S_{h,y}^2}{w_h}}{V + \sum_{h=1}^{H} N_h S_{h,y}^2}$

- **Margem de erro** para o estimador $\hat{\theta}$ (approx. normal para dist. de $\hat{\theta}$)
  - **Absoluta**: $e = z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})} \Leftrightarrow V = \frac{e^2}{z^2_{\frac{\alpha}{2}}}$
  - **Relativa**: $r\overline{Y} = z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})} \Leftrightarrow V = \frac{r^2\overline{Y}^2}{z^2_{\frac{\alpha}{2}}}$

Aula passada

- Tamanho mínimo de amostra para **estimação da média** populacional

Alocação	AASc dentro dos estratos	AASs dentro dos estratos
AES_{un}	$n \geq \frac{H \sum_{h=1}^{H} W_h^2 Var_{h,y}}{V}$	$n \geq \frac{H \sum_{h=1}^{H} W_h^2 S_h^2}{V + \frac{1}{N} \sum_{h=1}^{H} W_h S_h^2}$
AES_{pr}	$n \geq \frac{\sum_{h=1}^{H} W_h Var_{h,y}}{V}$	$n \geq \frac{\sum_{h=1}^{H} W_h S_h^2}{V + \frac{1}{N} \sum_{h=1}^{H} W_h S_h^2}$
AES_{ne}	$n \geq \frac{\left(\sum_{h=1}^{H} W_h DP_{h,y}\right)^2}{V}$	$n \geq \frac{\left(\sum_{h=1}^{H} W_h S_{h,y}\right)^2}{V + \frac{1}{N} \sum_{h=1}^{H} W_h S_h^2}$
AES_{ot}	$n \geq \frac{\left(\sum_{h=1}^{H} W_h DP_{h,y} \sqrt{C_h}\right)\left(\sum_{h=1}^{H} W_h DP_{h,y} / \sqrt{C_h}\right)}{V}$	$n \geq \frac{\left(\sum_{h=1}^{H} W_h S_{h,y} \sqrt{C_h}\right)\left(\sum_{h=1}^{H} W_h S_{h,y} / \sqrt{C_h}\right)}{V + \frac{1}{N} \sum_{h=1}^{H} W_h S_h^2}$

Aula passada 💾

Exemplo 🏋️

Exemplo 7 da Apostila (da Profa. Vanessa):

Suponha que os restaurantes em uma cidade foram divididos em 3 estratos, de acordo com a zona de localização: $A\ (N_1 = 600)$, $B\ (N_2 = 300)$ e $C\ (N_3 = 100)$. Queremos estimar o número médio de clientes por dia. Os valores do desvio padrão dos estratos são: 20, 30 e 50 clientes, respectivamente. Determinar o tamanho de amostra pra estimar a média de clientes por dia com um erro máximo absoluto de 3 clientes e com 99,73% de confiança (isto é, z = 3). Considere que será feita uma **AASs** em cada estrato.

Alguém tentou❓❓❓ Dúvidas ❓

Exercícios para entregar

Aula de Hoje

- São **três** exercícios para entregar. 🏃
    - Pode ser feito **à mão** ou **em códigos**, de qualquer forma serão **postados no moodle**.
    - Indicar **notações** e **fórmulas** utilizadas.
    - Mostrar **desenvolvimento**, **interpretação** e **conclusão**.
- Discutam as resoluções com os colegas, mas a **entrega** é **individual**.

Exercícios para entregar 1 💪

- Exercício 4.1 (Elementos de Amostragem)
  - Nos slides *Aula 06*, página 8, continuar ítens (c) e (d).
  - Interprete os resultados.

Exercícios para entregar 2 🏋️‍♀️

- Exercício 11.10 (Amostragem: Teoria e Prática Usando o R)

(Adaptado de @Scheaffer2011) Uma empresa tem suas divisões localizadas em três continentes distintos: América, Europa e Ásia. Deseja-se realizar uma pesquisa sobre um de seus produtos através de uma amostra de clientes a serem entrevistados por telefone a partir da divisão localizada na América. O custo das ligações é diferente para cada uma das divisões. A Tabela abaixo contém as informações do custo, em dólares, de cada ligação/entrevista para cada uma das divisões, além da variância das taxas de satisfação e o número total de clientes em cada estrato. Calcule o tamanho total da amostra a ser selecionada e a alocação apropriada para essa amostra, sabendo que se deseja que a variância da estimativa da média populacional seja $V_{AES}(\overline{y}_{AES}) \leq 0,1$. (assumindo **AASs** dentro dos estratos)

Estrato	N_h	$S^2_{h,y}$	C_h
América	112	2,25	9
Europa	68	3,24	25
Ásia	39	3,24	36

Exercícios para entregar 3 🏃‍♀️

- Exercício 11.7 (Amostragem: Teoria e Prática Usando o R)

As 2.120 lojas de uma certa localidade foram estratificadas pelo número de empregados (única variável relativa ao tamanho da empresa encontrada no cadastro) numa pesquisa para estimar o faturamento total. A Tabela abaixo contém as informações da variável número de empregados, x, utilizadas no planejamento da amostra e os resultados sobre o faturamento, y, das lojas por estrato, obtidos na coleta dos dados na amostra. O faturamento foi medido em 1.000 Reais. (assumindo **AASs** dentro dos estratos)

Estratos	N_h	$T_{h,x}$	$S^2_{h,x}$	$\bar{y}_h$ **(1.000 Reais)**	$\widehat{S}^2_{h,y}$
5-14	1.100	9.020	8,30	3	2,53
15-49	500	13.500	102,08	17	66,59
50-99	250	17.750	207,00	52	411,28
100-199	130	17.329	840,10	170	1.953,64
200-499	120	36.600	7.500,00	350	16.770,25
500 e mais	20	14.280	20.805,00	7.000	3.062.500,00

Exercícios para entregar 3 🏋️‍♀️

- (cont.) Exercício 11.7 (Amostragem: Teoria e Prática Usando o R)

a. Dimensione, utilizando os dados referentes a x, a amostra necessária para estimar o numero total de empregados com um erro máximo admissível de 2% e com um nível de confiança de 95%, supondo alocação de Neyman.

b. Suponha que a amostra alocada no item anterior tenha sido efetivamente selecionada fornecendo os resultados apresentados para y. Com estas informações estime o faturamento total, Y, e o coeficiente de variação desta estimativa.

Resoluções

Resolução Exercícios para entregar 1

```r
## dados do problema - Exercício 4.1 (Elementos de Amostragem)
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.
## media populacional
Ybarra <- sum( Nh * Ybarrah) / N        # media pop global
Ybarra
```

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

```r
## variancia populacional Var_y
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
```

```
## [1] 4.301073
```

Resolução Exercícios para entregar 1

```r
## b.
## alocacao proporcional
nhpr <- n * Nh / N                          # vetor de nh's na propor
nhpr
```

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

```r
## alocancao de neyman SEM reposicao dentro dos estratos
Sh <- sqrt(S2h)                             # desvios SHy dos estrato
nhneAESs <- n * (Nh * Sh) / (sum(Nh * Sh))  # vetor de nh's na de Ney
nhneAESs
```

```
## [1] 22.844026 23.819094 13.419060  9.791811 10.126008
```

```r
## alocacao de neyman COM reposicao dentro dos estratos
Varh <- (Nh - 1) * S2h / Nh                 # variancias dos estratos
DPh <- sqrt(Varh)                           # desvios padroes dos es
nhneAESc <- n * (Nh * DPh) / (sum(Nh * DPh))# vetor de nh's na de Ney
nhneAESc
```

Resolução Exercícios para entregar 1

```
## c.
## na AASc
Varybarra <- Vary / n                    # variancia de ybarra sob AASc
Varybarra
```

```
## [1] 0.05376341
```

```
## AESne sob AAS SEM reposicao dentro dos estratos
Wh <- Nh / N
VarybarraAESnes <- sum( Wh * Sh)^2 / n - sum( Wh * S2h) / N
VarybarraAESnes
```

```
## [1] 0.01532154
```

```
## AESne sob AAS COM reposicao dentro dos estratos
VarybarraAESnec <- sum( Wh * DPh)^2 / n
VarybarraAESnec
```

```
## [1] 0.01928409
```

```
## variancias reduzem com alocacao de neyman, vantagem sob ASSs dent
```

Resolução Exercícios para entregar 1

No item (a) calculamos $Var_y = 4.3010728$ e do item (b) temos $n = 80$.

(c)

- Sabemos que na **AASc** (ignorando os estratos) a **variância** do **estimador** da **média** é dada por

$$Var_{AASc}(\overline{y}) = \frac{Var_y}{n} = \frac{4.3010728}{80} = 0.0538.$$

- Já na **AESne** considerando **AASs** dentro dos estratos sabemos que

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

- E na **AESne** considerando **AASc** dentro dos estratos

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

Resolução Exercícios para entregar 1

```
## d.
## na AESpr sob AAS SEM reposicao dentro dos estratos
VarybarraAESprs <- sum(Wh * S2h) * ((1/n) - (1/N))
VarybarraAESprs
```

```
## [1] 0.01558885
```

```
## na AESpr sob AAS COM reposicao dentro dos estratos
VarybarraAESprc <- sum(Wh * Varh) / n
VarybarraAESprc
```

```
## [1] 0.019544
```

```
## variancias um pouco maiores que na alocacao ne neyman, novamente
```

Resolução Exercícios para entregar 1

(d)

- Para a variância na **AESpr** considerando **AASs** dentro dos estratos sabemos que

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

- E para a variância na **AESpr** considerando **AASc** dentro dos estratos

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

Resolução Exercícios para entregar 2

```r
## dados do problema - Exercício 11.10 (Amostragem: Teoria e Prática
H <- 3                              # no. de estratos
h <- 1:H                            # indice dos estratos
Nh <- c( 112, 68, 39)               # tamanho dos estratos
S2h <- c( 2.25, 3.24, 3.24)         # variancia do estrato
Ch <- c( 9, 25, 36)                 # custo de amostragem no estrato
N <- sum(Nh)                        # tamanho da populacao
V <- 0.1                            # variancia maxima

## calculo de n
Wh <- Nh / N          # peso do estrato h na pop.
Sh <- sqrt(S2h)       # variancia do estrato h
raizCh <- sqrt(Ch)    # raiz quadrada do custo no estrato h
num_part1 <- sum( Wh * Sh * raizCh)
num_part2 <- sum( Wh * Sh / raizCh)
denom <- V + sum( Wh * S2h) / N
n <- num_part1 * num_part2 / denom # tamanho da amostra sob AESot e /
n     # arredondar para cima, ceiling(n)
```

```
## [1] 26.26596
```

Resolução Exercícios para entregar 2

- Para calcular o tamanho total da amostra n, sob alocação ótima (*uma vez que o custo de observação das unidades difere de estrato para estrato*),
  - assumindo **AASs** dentro dos estratos,
  - e definindo a variância da estimativa da média populacional tal que não ultrapasse $V = 0,1$

$$
\begin{aligned}
Var_{AES_{ot}}(\overline{y}_{AES}) \leq 0,1 &\Leftrightarrow \frac{\left(\sum_{h=1}^{H} W_h S_{h,y}\sqrt{C_h}\right)\left(\sum_{h=1}^{H} W_h S_{h,y}/\sqrt{C_h}\right)}{0,1 + \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2} \leq n \\
&\Leftrightarrow \frac{7.0191781 \times 0.4209132}{0.1124826} \leq n \\
&\Leftrightarrow 26.2659638 \leq n.
\end{aligned}
$$

- Arredondaremos n para o inteiro mais próximo.
  - Então $n = 27$.

Resolução Exercícios para entregar 2

```
n <- ceiling(n)
nh <- n * (Wh * Sh / raizCh) / sum( Wh * Sh / raizCh) # tamanho da ar
```

- A alocação apropriada para essa amostra, assumindo **AESot** e **AASs** dentro dos estratos,

$$n_h = n \times \frac{N_h S_{h,y}}{\sum_{k=1}^{H} N_k S_{k,y}} = (16.4027, 7.1703, 3.427).$$

Sob alocacao ótima arredondar para o inteiro mais próximo nos estratos com menor custo, maior variabilidade, maior tamanho?

```
## arredondando todos para mais
ceiling(nh); sum( ceiling(nh))
```

```
## [1] 17  8  4
```

```
## [1] 29
```

```
## arredondando inteiro mais pro
round( nh); sum(round( nh))
```

```
## [1] 16  7  3
```

```
## [1] 26
```

Resolução Exercícios para entregar 3

```r
## dados do problema - Exercício 11.7 (Amostragem: Teoria e Prática (
H <- 6                              # no. de estratos
h <- 1:H                            # indice dos estratos
Nh <- c( 1100, 500, 250, 130, 120, 20)  # tamanho dos estratos
Thx <- c( 9020, 13500, 17750, 17329, 36600, 14280)
S2hx <- c( 8.30, 102.08, 207.00, 840.10, 7500.00, 20805.00)   # vari
ybarrah <- c( 3, 17, 52, 170, 350, 7000)                     # media
s2hy <- c(2.53, 66.59, 411.28, 1953.64, 16770.25, 3062500.00) # vari
N <- sum(Nh)                        # tamanho da populacao

## a.
## valor de V
r <- 0.02                              # erro relativo 2%
alfa <- 0.05                           # confianca (1 - alfa) = 95%
z_alfa_2 <- qnorm(1-alfa/2)            # usando aproximacao pela normal
Wh <- Nh / N                           # peso do estrato h na pop.
Tx <- sum(Thx)                         # total populacional de x
Xbarra <- Tx / N                       # media populacional de x
Shx <- sqrt(S2hx)                      # variancia de x do estrato h
```

Resolução Exercícios para entregar 3

```r
## calculo de n
## se fosse estimacao da media
V <- r^2 * Xbarra^2 / z_alfa_2^2 # variancia maxima - erro relativo p
num <- sum( Wh * Shx)^2          # num. formula n para media
denom <- V + sum( Wh * S2hx) / N # denom. formula n para media
n <- num / denom  # tamanho da amostra sob AESot e AASs dentro para a
n                 # arredondar para cima, ceiling(n)
```

```
## [1] 301.5521
```

```r
## na estimacao do total
V <- r^2 * Tx^2 / z_alfa_2^2 # variancia maxima - erro relativo para
num <- sum( Nh * Shx)^2      # num. formula n para total
denom <- V + sum( Nh * S2hx) # denom. formula n para total
n <- num / denom  # tamanho da amostra sob AESot e AASs dentro para o
n                 # arredondar para cima, ceiling(n)
```

```
## [1] 301.5521
```

```r
## usando formula de n para media com V adequado para media é igual a
## n usando a formula para o total e V adequado para total
```

Resolução Exercícios para entregar 3

(a)

- Utilizando os dados referentes a x, o tamanho de amostra necessária para estimar o número total de empregados com um erro máximo admissível de 2% e com um nível de confiança de 95%, supondo alocação de Neyman, na **AASs** dentro do estratos

  - Se fossemos usar a fórmula para a estimação da média,

$$\overline{X} = \frac{T_x}{N} = \sum_{h=1}^{H} T_{h,x}/N = (9020 + 13500 + 17750 + 17329 + 36600 + 14280)/N = 51.1693396.$$

- A variância mínima é dada por $V = \frac{r^2 \overline{X}^2}{z_{\frac{\alpha}{2}}^2}$, para estimação da média quanto do total(?)

$$n \geq \frac{\left(\sum_{h=1}^{H} W_h S_{h,x}\right)^2}{V + \frac{1}{N}\sum_{h=1}^{H} W_h S_{h,x}^2} = 301.5520596.$$

- Então, $n = 302$.

Resolução Exercícios para entregar 3

- E na **estimação do total**???

$$n \geq \frac{\sum_{h=1}^{H} \frac{N_h^2 S_{h,y}^2}{w_h}}{V_T + \sum_{h=1}^{H} N_h S_{h,y}^2}$$

No caso da estimação do total V_T segue sendo uma variância máxima, porém agora será máxima para a variância do estimador do total

$$rN\overline{Y} = z_{\frac{\alpha}{2}} \sqrt{Var(\overline{T}_{AES})} \Leftrightarrow V_T = \frac{r^2 N^2 \overline{Y}^2}{z_{\frac{\alpha}{2}}^2}.$$

- O **tamanho da amostra** n é o **mesmo** na estimação da **média** e **total**,
  - se usar erro relativo $r\overline{Y}$ para definir V e usar $V \geq Var_{AES}(\overline{y})$;
  - ou com erro relativo $rN\overline{Y}$ para definir V_T e usar $V_T \geq Var_{AES}(\widehat{T})$.

Resolução Exercícios para entregar 3

```
## b.
## alocacao de neyman
nh <- ceiling(n * Shx * Nh / sum(Shx * Nh))       # arredondando para
nh
```

```
## [1]  34  53  38  40 109  31
```

```
## maior que a população no estrato 6???
Nh
```

```
## [1] 1100  500  250  130  120   20
```

```
## estimativa do faturamento total sob AASs dentro
ty <- sum(Nh * ybarrah)                             # estimativa do total
ty
```

```
## [1] 228900
```

```
## variancia estimada do estimador do total
s2ty <- sum( Nh^2 * (1/nh - 1/Nh) *  s2hy)   # estimativa da varianc
cvty <- sqrt( s2ty) / ty                      # coeficiente de variaca
```

Resolução Exercícios para entregar 3

(b)

- Supondo que a amostra alocada no item (a) tenha sido efetivamente selecionada fornecendo os resultados apresentados para y. a estimativa do faturamento total, Ty,

$$\widehat{T}_{y,AES} = \sum_{h=1}^{H} \widehat{T}_{h,y} = \sum_{h=1}^{H} N_h \bar{y}_h$$

- E o coeficiente de variação desta estimativa, ...
  - Para o cálculo de

$$\widehat{Var}_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$$

- temos $n_h > N_h$(❓)

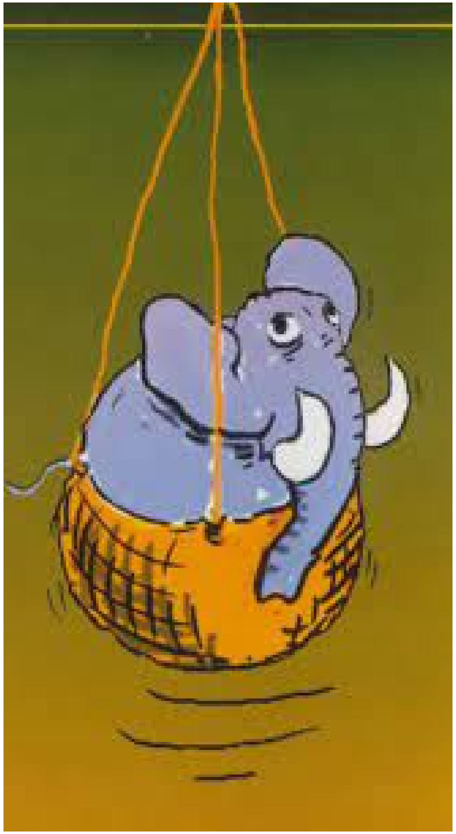
Para casa 🏠

- Continuar os Exercícios e Entregar.
- Continuar exercícios 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

Referências

- Amostragem: Teoria e Prática Usando o R