MAT02036 - Amostragem 2

Aula 10 - Amostragem Estratificada - Avaliação Parcial 1

Markus Stein

Departamento de Estatística, IME/UFRGS

2022/2

Housekeeping

- A avaliação é composta por **cinco** exercícios. 
    - 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**.
    - colaboração é importante, mas cópia é proibido.

Extra: (1.0 ponto) para quem entregou o exercício da aula 08.

Boa Avaliação!

Avaliação Parcial 1

Exercício 1 

(2,0 pontos) Revisar exercícios 1 e 2 com resolução nos slides 'Aula 08'.

- Quem já entregou, conferir seus resultados com a resolução atualizada nos slides da `Aula 08`.
  - Alguma discordância com a resolução?
- Quem fez parcialmente e/ou não entregou, ...

Exercício 2 

(2,0 pontos) Exercício 3 dos slides 'Aula 08' modificado.

a. Conferir cálculos da letra (a) do exercício original nos slides da `Aula 08`.

b. A alocação apresenta $n_h > N_h$? Se sim, como realocar?

c. Apresente uma estimativa do faturamento total T_y por ponto e por intervalo, de acordo com o realocamento em (b). Interprete.

(bônus): No item (a), mostrar que tamanho da amostra para média é equivalente ao tamanho para o total, usando V e a fórmula de n adequados para cada caso.

Exercício 3 

(2,0 pontos) Banco de dados `Lucy`. (arquivo `Aula_AE_xxx.R`)

Assuma que o banco de dados `Lucy` represente uma população.

a. Apresente o tamanho na população por estrato, os parâmetros média e variância populacionais da variável `Income`, por estrato e globias, usando `Zone` como variável estratificadora. Repita para a variável `Level` como estratificadora.

b. Supondo **AASc** de 30 empresas em cada estrato, calcule a variância do estimador da média de `Income`, comparando o desempenho de `Zone` e `Level` como variáveis estratificadoras.

c. Qual a melhor variável para estratificação? Justifique, por exemplo, calculando efeito de planejamento EPA $(deff)$ em cada caso.

d. Selecione uma amostra com a alocação definida no item (b) e obtenha uma estimativa pontual e intervalar para a média de `Income`. Interprete.

(bônus): Adicionar comandos para obtet total de `Income`.

(bônus 2): Adicionar comandos para obter proporção de `SPAM`.

Exercício 4 💪 - *Assuma* **AASc** *dentro dos estratos*

(2,0 pontos) Exercício 4.4 (elementos de amostragem)

4.4 Planejou-se uma amostragem estratificada com reposição para estimar a porcentagem de famílias tendo conta em caderneta de poupança e também da quantidade investida. De uma pesquisa passada, têm-se estimativas para as proporções P_h e para os desvios padrões das quantidades investidas, σ_h, conforme descrito na tabela abaixo.

h	W_h	P_h	σ_h
1	0,6	0,20	9
2	0,3	0,40	18
3	0,1	0,60	52

Calcule os menores n e n_h que satifaçam, com custo constante:

a. A proporção populacional dever ser estimada com erro padrão igual a 0,02;

b. A quantidade média investida deve ser estimada com erro padrão igual a R$ 2,00.

Qual dos tamanhos, em (a) ou em (b), você usaria na pesquisa? Por quê?

Exercício 5 🏋 - *Assuma* **AASc** *dentro dos estratos*

(2.0 pontos) Exercício 4.20 (elementos de amostragem)

4.20 Uma cadeia de lojas está interessada em estimar, dentro das contas a receber, a proporção das que dificilmente serão recebidas. Para reduzir o custo da amostragem, usou-se AE com cada loja num estrato. Os dados obtidos foram os seguintes:

h	N_h	n_h	$\hat{P}_h$
1	60	15	0,30
2	40	10	0,20
3	100	20	0,40
4	30	6	0,10

onde N_h é o número de contas a receber, n_h é o tamanho da amostra e $\hat{P}_h$ é a proporção de contas problemáticas. Dê uma estimativa para a proporção total de quatro lojas e um intervalo de confiança de 95% para a mesma.

Para casa 

- Continuar os Exercícios e Entregar.

Próxima aula 

- Acompanhar o material no moodle.

Boas Festas! 🎅

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)

Solução:

Exercício 1

(2,0 pontos) Revisar exercícios 1 e 2 com resolução nos slides 'Aula 08'.

Exercício 4.1 (Elementos de Amostragem), Itens (c) e (d).

```
## 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
```

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

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

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

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

```
## [1] 22.901909 23.859484 13.419324  9.737872 10.081411
```

Exercício 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
```

Exercício 1

(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

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

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

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

As Variâncias nas estratégias de alocação de Neyman são consideravelmente menores. A variância do estimador da média no plano **AASc** ignorando os estratos é cerca de 3,5 vezez a variância assumindo **AESne** sob **AASs**. Se **AASc** dentros dos estratos, esse número cai para 2,8.

Exercício 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
```

Exercício 1

(d)

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

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

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

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

Usando alocação proporcional, a redução nas variâncias é muito similar a da alocação de Neyman, com uma pequena vantagem para a última.

Exercício 1

Ex. 11.10 (Amostragem: Teoria e Prática Usando o R)

```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 A
# n     # arredondar para cima, ceiling(n)
```

Exercício 1

- 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 $V = 0,1$

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

- Arredondaremos n para o número inteiro maior e mais próximo, então precisamos de no mínimo $n = 27$ observações para garantir que a variância do estimador da média não ultrapasse $V = 0,1$.

Exercício 1

```
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
```

Exercício 2

(2,0 pontos) Exercício 3 dos slides 'Aula 08' modificado.

(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

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

Denotamos V_T um valor máximo para a variância do estimador do total, então

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

Temos que

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

então $n = 302$.

Exercício 2

(b) Conforme os cálculos da letra (a) do exercício original nos slides da `Aula 08`, a alocação apresenta $n_6 > N_6$.

- Uma solução é observar todas as unidades do estrato em que $n_H \geq N_H$.
- Definindo $n_6 = N_6$, restam $N^* = N - N_6 = 2120 - 20 = 2100$ unidades da população para serem selecionadas e distribuídas nos $H^* = 5$.
- O tamanho restante da amostra é $n^* = n - N_6 = 302 - 20 = 282$.
- Sob alocação de Neyman, agora temos $n_h^* = n^* \frac{S_h N_h}{\sum_k S_k N_k}$, $h, k = 1, \ldots, 5$.
  $N_h S_{h,x} = (1100 \times 2.88; 500 \times 10.1; 250 \times 14.39; 130 \times 28.98; 120 \times 86.6)$ e $\sum_{k=1}^{5} N_h S_{h,x} = 28862.75$. Assim,

```
nh_novo <- n_novo * (Nh * Shx)[1:5] / sum( (Nh * Shx)[1:5])
(nh <- c( nh_novo, Nh[6]))      # nova alocacao incluindo o ultimo es
```

```
## [1]  34.40137  54.83835  39.04534  40.90278 112.81216  20.00000
```

Exercício 2

(c) De acordo com o realocamento em (b), uma estimativa pontual do faturamento total é dada por

$t_y = \widehat{T}_y = \sum_{h=1}^{6} \widehat{T}_{h,y} = \sum_{h=1}^{6} N_h \bar{y}_h = 11003 + 50017 + 25052 + 130170 + 120350 + 207000.$

```
ty <- sum(Nh * ybarrah)                    # estimativa do total
```

2.289\times 10^{5} milhares de reais.

Uma estimativa intervalar para o faturamento total é dada por (se $n = \sum_{h=1}^{H} n_h$ "for grande"),

$$IC(T_y; 0, 95) = \left[t_y \pm z_{0,05/2} \sqrt{\widehat{Var}_{AES}\left(t_y\right)} \right]$$

Assumindo (o coeficiente de) confiança $1 - \alpha = 95\%$, por exemplo, temos o valor da distribuição normal padrão que deixa área 0,025 à sua direita dado por $z_{0,025} = 1.959964$.

Exercício 2

Sob *AASs* dentro dos estratos, temos os estimadores da variância do estimador do total $s^2_{h,t_y} = Var_{AES}\left(\widehat{T}_h\right) = N_h^2 Var_{AES}\left(\bar{y}_h\right)$.

A variância global é dada por $s^2_{t_y} = \widehat{Var}\left(t_y\right) = \sum_{h=1}^{6} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s^2_{h,t_y}$

```
s2ty <- sum( Nh^2 * (1/nh - 1/Nh) *  s2hy)   # estimativa da varianc
s2ty
```

```
## [1] 1593445
```

Assim, temos o erro absoluto, $e = z_{0,05} * \sqrt{s^2_{t_y}}$

```
e <- -qnorm(0.025) * sqrt(s2ty)   # erro absoluto
e
```

```
## [1] 2474.096
```

Exercício 2

O intervalo ao nível 95% para T_y é dado por

```
IC <- ty + c(-1, 1) * e          # intervalo de confianca para o tota
IC
```

```
## [1] 226425.9 231374.1
```

O intervalo de 226425.9 milhares de reais a 231374.1 milhares de reais deve conter o verdadeiro faturamento total dos estabelecimentos, com uma confiança de 95%. Ou ainda, a cada 100 amostras observadas sob o mesmo plano, espera-se que 95% dos intervalos construídos com base nessas amostras conterão o verdadeiro faturamento total dos estabelecimentos.

Exercício 2

(*Bônus*) No item (a), mostrar que tamanho da amostra para média é equivalente ao tamanho para o total, usando V e a fórmula de n adequados para cada caso.

- 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 $V_T \geq Var_{AES}(\widehat{T})$.

A variância mínima dos estimador da média é dada por $V = \frac{r^2\overline{X}^2}{z^2_{\frac{\alpha}{2}}}$.

Exercício 3

(2,0 pontos) Banco de dados `Lucy`. (arquivo `Aula_AE_xxx.R`)

a. Usando `Zone` como variável estratificadora, temos o tamanho da população, bem como a média e variância populacionais da variável `Income`, por estrato e globais,

Parâmetros por estrato

```
## parametros por estrato
Ybarrah <- tapply( Lucy$Income, Lucy$Zone, mean)      # medias - ou a
Varh <- tapply( Lucy$Income, Lucy$Zone, varpop)       # Variancias
cbind( h = 1:H, Nh = Nh, Media = Ybarrah, Variancia =  Varh)
```

```
##   h  Nh    Media Variancia
## A 1 307 652.2834  78955.58
## B 2 727 320.7469  51586.62
## C 3 974 331.0195  23561.77
## D 4 223 684.9821  43597.21
## E 5 165 767.3879  58406.46
```

Exercício 3

Parâmetros globais

```
## parametros globais
Ybarra <- mean( Lucy$Income)  # media
# sum( Wh * Ybarrah)          # forma equivalente para caluclar media
Ybarra
```

```
## [1] 432.0605
```

```
Var <-  varpop(Lucy$Income)   # Variancia
Var
```

```
## [1] 71248.14
```

```
# Vard <- sum( Wh * Varh)                    # Variância dentro
# Vard
# Vare <- sum( Wh * (Ybarrah - Ybarra)^2) # Variância entre
# Vare
# Vare + Vard                                # equivalente ao calculo de
```

Exercício 3

Usando `Level` como variável estratificadora. temos

Parâmetros por estrato

```
## parametros por estrato
Ybarrah2 <- tapply( Lucy$Income, Lucy$Level, mean)      # medias
Varh2 <- tapply( Lucy$Income, Lucy$Level, varpop)       # Variancias
cbind( h = 1:H2, Nh = Nh2, Media = Ybarrah2, Variancia =  Varh2)
```

```
##        h   Nh     Media Variancia
## Big    1   83 1249.4699  63622.27
## Medium 2  737  661.2632  16060.08
## Small  3 1576  281.8274  15132.44
```

Os Parâmetros globais devem ser os mesmos

```
## globais
Ybarra <- mean( Lucy$Income)  # media
Ybarra
```

```
## [1] 432.0605
```

Exercício 3

b. Supondo **AASc** de 30 empresas em cada estrato, a variância do estimador da média de `Income`, utilizando `Zone` como variável estratificadora,

```
# Supondo AASc de 30 empresas em cada estrato, vamos calcular a varia
nh <- rep(30, H)
Varybarrah <- Varh/nh
Varybarrah
```

```
##         A         B         C         D         E
## 2631.8528 1719.5539  785.3922 1453.2403 1946.8819
```

```
Varybarra <- sum( Wh^2 * Varybarrah)
Varybarra
```

```
## [1] 353.1272
```

```
## Ignorando os estratos, a variância do estimador da média se AASc
Var/(30*H)
```

```
## [1] 474.9876
```

Exercício 3

Para `Level` como variável estratificadora,

```
# Supondo AASc de 30 empresas em cada estrato, vamos calcular a varia
nh2 <- rep(30, H2)
Varybarrah2 <- Varh2/nh2
Varybarrah2
```

```
##       Big    Medium     Small
## 2120.7424  535.3359  504.4148
```

```
Varybarra2 <- sum( Wh2^2 * Varybarrah2)
Varybarra2
```

```
## [1] 271.432
```

```
## Ignorando os estratos, a variância do estimador da média se AASc
Var/90
```

```
## [1] 791.646
```

Exercício 3

c. Se olharmos diretamente para as variâncias dos estimadores, usando $n_h = 30$ paratodos os estratos, para ambas as variáveis estratificadoras, temos

Variável estratificadora	$Var_{AES}\left(\bar{y}\right)$	n
`Zone`	353.1272237	150
`Level`	271.4319711	90

A comparação das variâncias pode não parecer justa à primeira vista, pois fixando $n_h = 30$ sendo que a variável `Zone` possui $H = 5$ estratos e `Level` $H = 3$. Mas ainda, usando $n_h = 50$ para `Level` a $Var_{AES}\left(\bar{y}\right)$ diminuiria.

Olhando para os **EPA**s para ambas as variáveis estratificadoras temos

Variável estratificadora	$Var_{AASc}\left(\bar{y}\right) = \frac{Var_y}{n}$	$EPA = \frac{Var_{AES}(\bar{y})}{Var_{AASc}(\bar{y})}$
`Zone`	$\frac{7.1248141 \times 10^4}{150} = 474.9876077$	0.7434451
`Level`	$\frac{7.1248141 \times 10^4}{90} = 791.6460129$	0.3428704

Exercício 3

d. Usando a variável `Level` como estratificadora, selecionamos uma amostra de tamanho $n = 90$ de igual tamanho entre os $H = 3$ estratos.

```
set.seed(02036)
nh <- 30
s <- tapply( Lucy$ID, Lucy$Level, sample, size=nh)              # IDs
Lucy_amostra <- Lucy[Lucy$ID %in% unlist(s), c("Level","Income")]
```

Ou podemos usar a função... do pacote... .

Com a amostra observada, calculamos os valores por estrato

```
## estimativas por estrato
ybarrah <- tapply( Lucy_amostra$Income, Lucy_amostra$Level, mean)
varh <- tapply( Lucy_amostra$Income, Lucy_amostra$Level, varpop)
cbind( h = 1:H2, Nh = Nh2, nh = nh, ybarra = ybarrah, varh =  varh)
```

```
##        h   Nh nh   ybarra     varh
## Big    1   83 30 1254.367 88396.30
## Medium 2  737 30  659.700 18700.74
## Small  3 1576 30  263.300 16775.41
```

Exercício 3

Uma estimativa do lucro médio das companhias, no particular ano fiscal, é aproximadamente

$$\begin{aligned}\overline{y} &= \sum_{h=1}^{3} W_h \overline{y}_h \\ &= 0.03 \times 1254.37 + 0.31 \times 659.7 + 0.66 \times 263.3 \\ &= 419.56 \text{reais.}\end{aligned}$$

Utilizando **AASc** dentro dos estratos, a **estimativa da variância da média amostral** é dada por

$$\begin{aligned}Var_{AES}\left(\overline{y}\right) &= \frac{1}{k}\sum_{h=1}^{3} W_h^2 Var_h \\ &= \frac{1}{30}(0 \times 63622.27 + 0.09 \times 16060.08 + 0.43 \times 15132.44) \\ &= 304.45.\end{aligned}$$

Exercício 3

Por fim, uma estimativa intervalar para o lucro médio das companhias, com 95% de confiança (se $n = \sum_{h=1}^{H} n_h$ "for grande") é dado por

$$IC(\overline{y}; 0,95) = \left[\overline{y} \pm z_{0,05/2}\sqrt{\widehat{Var}_{AES}\left(\overline{y}\right)}\right]$$

Assumindo (o coeficiente de) confiança $1 - \alpha = 95\%$, por exemplo, temos o valor da distribuição normal padrão que deixa área 0,025 à sua direita dado por $z_{0,025} = 1.959964$. Assim, o erro absoluto, $e = z_{0,05} * \sqrt{s^2_{\overline{y}}}$ é aproximadamente

```
e <- -qnorm(0.025) * sqrt(var_ybarra)   # erro absoluto
e
```

```
## [1] 34.19818
```

E o intervalo ao nível 95% para $\overline{y}$ é dado por

```
IC <- ybarra + c(-1, 1) * e          # intervalo de confianca para a
IC
```

Exercício 3

O intervalo de 385.36 dólares a 453.76 dólares deve conter o verdadeiro lucro médio da companhias, no ano fiscal estudado, com uma confiança de 95%. Ou ainda, a cada 100 amostras observadas sob o mesmo plano, espera-se que em 95 dos intervalos construídos com base nessas amostras conterão o verdadeiro lucro médio das companhias.

(bônus): Adicionar comandos para obter total de `Income`.

(bônus 2): Adicionar comandos para obter proporção de `SPAM`.

Exercício 4

(2,0 pontos) Exercício 4.4 (elementos de amostragem)

a. Para estimar uma proporção de famílias que possume conta em caderneta de poupança, com erro padrão da estimativa de no máximo 0,02, definimos $V_P \geq Var_{AES}\left(\widehat{P}_{AES}\right)$.

Lembrando erro padrão: $EP\left(\widehat{P}\right) = \sqrt{Var\left(\widehat{P}\right)} = DP\left(\widehat{P}_{AES}\right)$.

Assumindo **AASc** dentro dos estrtos, e custo de amostragem constante, o tamanho amostral mínimo para assegurarmos uma variância da estimativa da proporção menor ou igual a $V_P = \left[EP\left(\widehat{P}_{AES}\right)\right]^2$ é dado por

$$n \geq \frac{\left(\sum_{h=1}^{H} W_h \sqrt{P_h(1-P_h)}\right)^2}{V_P} = \frac{\left(\sum_{h=1}^{H} W_h \sqrt{P_h(1-P_h)}\right)^2}{\left[EP\left(\widehat{P}_{AES}\right)\right]^2}$$

$$= \frac{0.6 \times 0.4 + 0.3 \times 0.49 + 0.1 \times 0.49}{(0,02)^2} = 475.15$$

Exercício 4

A partição ótima com custo constante, partição de Neyman, nesse caso (**AASc** dentro) pode ser dada por

$$n_h = n \times \frac{w_h \times \sqrt{P_h(1-P_h)}}{\sum_k w_k \times \sqrt{P_k(1-P_k)}} = (261.58, 160.18, 53.39)\,.$$

b. Para estimar a quantidade média, com custo constante, sabemos que

$$n \geq \frac{\left(\sum_{h=1}^{H} W_h DP_{h,y}\right)^2}{V}.$$

Fixando o erro padrão máximo da estimativa da média em 2 *reais*, temos $V = 4 reais^2$, assim

$$n \geq \frac{0.6 \times 9 + 0.3 \times 18 + 0.1 \times 52}{2^2} = 64.$$

E a partição dada por

$$n_h = n \times \frac{w_h \times DP_h}{\sum_k w_k \times DP_k} = (21.6, 21.6, 20.8)$$

Interpretação e conclusão: Não havendo restrição de orçamento na pesquisa

Exercício 5

Uma estimativa pontual da proporção de contas problemáticas $\widehat{P}_{AES}$ é dada por

$$\begin{aligned}\widehat{P} &= \sum_{h=1}^{4} W_h p_h \\ &= 0.26 \times 0.3 + 0.17 \times 0.2 + 0.43 \times 0.4 + 0.13 \times 0.1 \\ &= 0.3.\end{aligned}$$

Ou seja, estimamos que a proporção de contas problemáticas da rede de lojas seja aproximadamente 30%.

Utilizando **AASc** dentro dos estratos, a **estimativa da variância da proporção amostral** é dada por

$$\begin{aligned}Var_{AES}\left(p\right) &= \frac{1}{k}\sum_{h=1}^{4} W_h^2 p_h(1-p_h) \\ &= \frac{1}{15,10,20,6}(0.07 \times 0.21 + 0.03 \times 0.16 + 0.19 \times 0.24 + 0.02 \times 0.09) \\ &= 0.0043.\end{aligned}$$

Exercício 5

Por fim, uma estimativa intervalar para a proporção de contas problemáticas, com 95% de confiança (se $n = \sum_{h=1}^{H} n_h$ "for grande") é dado por

$$IC(P; 0,95) = \left[p \pm z_{0,05/2} \sqrt{\widehat{Var}_{AES}\,(p)} \right]$$

Assumindo (o coeficiente de) confiança $1 - \alpha = 95\%$, por exemplo, temos o valor da distribuição normal padrão que deixa área 0,025 à sua direita dado por $z_{0,025} = 1.959964$. Assim, o erro absoluto, $e = z_{0,05} * \sqrt{s_p^2}$ é aproximadamente

```
e <- -qnorm(0.025) * sqrt(var_p)   # erro absoluto
e
```

```
## [1] 0.1278124
```

E o intervalo ao nível 95% para $\bar{y}$ é dado por

```
IC <- p + c(-1, 1) * e          # intervalo de confianca para a media
IC
```

```
## [1] 0.1721876 0.4278124
```

Exercício 5

O intervalo de 0.17 a 0.43 deve conter a verdadeira proporção de contas problemáticas dessa rede, com uma confiança de 95%. Ou ainda, a cada 100 amostras observadas sob o mesmo plano, espera-se que em 95 dos intervalos construídos com base nessas amostras conterão a verdadeira proporção de contas problemáticas.

Pontuação e Comentários

Pontuação

Avaliação parcial 1

Exercício 1: (2,0 pontos)

- Ex. 4.1 (Elementos de Amostragem),
  a. (0,8 pontos) Itens (c) e (d).
  b. (0,2 pontos) Interprete os resultados.

- Ex. 11.10 (Amostragem: Teoria e Prática Usando o R)
  a. (0,1 pontos) Informações do enunciado.
  b. (0,2 pontos) Indicar fórmulas.
  c. (0,6 pontos) Cálculo de n e n_h.
  d. (0,1 pontos) Interpretação.

Exercício 2: (2,0 pontos)

a. (0,6 pontos) Cálculo n e nh conferir.
b. (0,7 pontos) Comentários e realocação.
c. (0,7 pontos) Estimação pontual, por intervalo e interpretação.
d. (*0,5 pontos bônus)

Pontuação

Avaliação parcial 1

Exercício 3: (2,0 pontos)

a. (0,4 pontos) Parâmetros por `Zone` e `Level`.
b. (0,6 pontos) Variância do estimador da média por `Zone` e `Level`.
c. (0,2 pontos) Comparação.
d. (0,8 pontos) Seleção da amostra, estimação e interpretações.
e. (*0,2 pontos bônus1)*
*f. (*0,3 pontos bônus2)

Exercício 4: (2,0 pontos)

a. (0,9 pontos) Cálculos e Interpretação
b. (0,9 pontos) Cálculos e Interpretação
c. (0,2 pontos) Conclusão

Exercício 5: (2,0 pontos)

a. (1,0 pontos) Estimativa pontual e interpretação
b. (1,0 pontos) Estimativa intervalo e interpretação

Pontuação

Atividade Aula 08

Exercício 1:

- a. Nos slides *Aula 06*, página 8, continuar ítens (c) e (d).
- Interprete os resultados.
- Arredondamento.

Exercício 2: Exercício 11.10 (Amostragem: Teoria e Prática Usando o R)

- Informações do enunciado.
- Indicar fórmula.
- Cálculo de n.
- Conclusáo... "tamanho mínimo"

Exercício 3: Exercício 11.7 (Amostragem: Teoria e Prática Usando o R)

- a. Expressões, cálculo e interpretação.
- b. Expressões, cálculo e interpretação.

Comentarios gerais

- Enviar o output `.pdf`, com imagens da resolução `à mão` ou os resultados da versão `em códigos`
- Se os códigos estão em arquivo `.R`, enviar fórmulas em separado
- Se enviar código com erro deixar claro onde está o erro... o ideal seria deixar as linhas com erro comentadas
- Colocar nome nos arquivos, nos códigos e outputs também.
- Fórmulas/expressões devem aparecer, seja no formato `à mão` ou `em códigos`... se enviar somente pdf... enviar arquivo de codigos tambem
- Interpretações... Pontual... qual o parâmetro em termos do problema, unidade de medida... ou IC parametro, unidade de medida, confianca... Resultado e Conclusão: O IC(), com alfa; Portanto, com 95%...
  - " estimamos que a média esteja entre 406,16 e 167,84 com 95% de confiança." ... "o total esta em ic de 95% de confiança entre 187184.3 e 192095.7" está correto:
  - suposições para ICs... Fórmulas... códigos... se pegamos de outros, cuidado para não deixar código desnecessário... manter organizados os códigos... Entrega... relatorio bem apresentado... "Marcia falou sobre relatorio e markdown"... capricho com relatorios Se nao fez a atividade "Aula 08" conferiu os codigos... odnde estao... tua versao dos codigos... semente nao aleatoria, nos caso de selecionar amostra cada um poderia ter uma amostra diferente...

Comentarios gerais

- Para evitar o `for()`
  - calcular somas de vetores... x <- 1; for{x <- x + 1}
  - calcular medidas agregadas `tapply()` faz o mesmo, ou `aggregate()`.
- Arredondamentos: n `ceiling()` e n_h `round()`
- Cálculo do deff incorreto (usando amostra???), selecionando amostra e usando `svymean`. (-0,2 pontos)