

MAT02036 - Amostragem 2

Aula 18 - Amostragem por Conglomerados - Avaliação Parcial 2

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

Exercícios e Lab

Utilizaremos o banco de dados Lucy (com informações ao nível individual) para: a. calcular os parâmetros e selecionar amostras b. calcular o coeficiente de correlação intraclass c. estimação e tamanho da amostra, IC

Parâmetros

Arquivo parametros e sorteio na AC1.R

Estimação, tamanho amostra e IC

Arquivo estimacao e tamanho de amostra AC1.R

CCI

Arquivos exemplo_pg31_apostila.R e exemplo_pg35_apostila.R

Avaliação Parcial 2

Avaliação Parcial 2

Instruções

- Responda individualmente os itens na caixa de texto, ou anexe um arquivo .pdf com:
 - Desenvolvimento e expressões (e códigos, se forem utilizados)
 - Resposta e interpretação.

Obs. 1: Na caixa de texto é possível colar figuras com desenvolvimentos, expressões (e códigos).

Obs. 2: O documento .pdf com desenvolvimento e expressões (códigos, se for o caso), pode ser único para todas as questões, nesse caso indicar na caixa de texto das questões.

Boa avaliação!

Avaliação Parcial 2

Questão 1

Considere uma população com $N = 8$ indivíduos, onde

$$Y = (9, 10, 11, 17, 20, 31, 32, 30).$$

a. Seja a divisão A desta população:

$$U_A = (C_1, C_2) = ((9, 10, 11, 17), (20, 31, 32, 30)).$$

Calcule o coeficiente de correlação intraclasse e o interprete. Qual é o menor valor que o coeficiente pode assumir nesse caso?

b. Considere agora a divisão B :

$$U_B = (C_1, C_2) = ((10, 20, 30, 11), (32, 9, 17, 31)).$$

Calcule o coeficiente de correlação intraclasse. Compare os resultados das duas divisões.

c. Na divisão A você recomendaria utilizar um plano **AC1S** ou **AAS**? E na divisão B ? Justifique.

Avaliação Parcial 2

Questão 2

Uma empresa de táxis deseja estudar a situação dos pneus dos veículos da sua frota, que é composta por 175 táxis. Para tanto, uma amostra de 10 táxis foi selecionada com reposição e, para cada um, se avaliou o número de pneus (dentre os 4 pneus em uso) que estavam fora de condições de segurança. Os resultados obtidos foram:

2, 4, 0, 1, 2, 0, 4, 1, 3, 1

- Estime a proporção de pneus da frota fora de condições pontualmente e por *IC* 95%.
- Usando esses resultados como um estudo piloto, qual seria o número de táxis necessário para obter uma estimativa da proporção de pneus fora das condições, com um erro absoluto de 2,5% e 95% de confiança? Considere **AC1s** com reposição.

Avaliação Parcial 2

Questão 3

Considere o banco de dados `agpop` do pacote `SDaA` do R. Após instalar o pacote, ao executar os comandos abaixo o banco de dados será carregado e poderá ser utilizado. Considere que os dados se referem a população de distritos dos EUA.

```
library(SDaA)
data(agpop)
```

Responda:

- Usando o seu cartão UFRGS como semente aleatória (`set.seed(XXXXXXX)`, onde `XXXXXXX` é o número do seu cartão), sorteie uma **AC1S** de 15 estados (variável `state`) sem reposição.
- A partir da amostra sorteada, obtenha e apresente a estimativa pontual e por *IC* 95% da média da variável `largef92`. Interprete os resultados.
- Produza dois gráficos que descrevam a variável `largef92`: um na população, outro na **AC1S** sem reposição sorteada. Comente sobre as diferenças encontradas.

Solução

Avaliação Parcial 2

Questão 1 - solução

Do enunciado temos:

- $N = 8$ unidades elementares na população,
- $M = 2$ conglomerados.

```
## exercicio 1
## dados do problema
Y <- c(9,10,11,17,20,31,32,30)      # vetor pop. de valores Y
CA <- c(1,1,1,1,2,2,2,2)           # indices cluster A
CB <- c(2,1,1,2,1,2,2,1)           # indices cluster B
```

a) O coeficiente de correlação intraclasse é dada pelo coeficiente de correlação linear de Pearson dos dos possíveis pares dentro dos conglomerados (Y_1', Y_2') ,

$$\rho = \frac{Cov(Y_1', Y_2')}{\sqrt{Var(Y_1')Var(Y_2')}}.$$

- Y_1' significa possíveis valores da primeira posição do par
- Y_2' significa possíveis valores da segunda posição do par.

Avaliação Parcial 2

Questão 1 - solução

Adaptando o código da Aula 14:

```
library(gtools)

## divisao A
Ni <- mean(table(CA))          # tamanho de cada conglomerado,
## cria os pares
YlinhaCA1 = permutations( length( Y[CA == 1]), 2, Y[CA == 1] ,set=F)
YlinhaCA2 = permutations( length( Y[CA == 2]), 2, Y[CA == 2] ,set=F)
YlinhaCA = rbind( YlinhaCA1, YlinhaCA2)  # pares de toda a pop.
rhoCA = cor( YlinhaCA[,1], YlinhaCA[,2]) # coef. corr. intraclasses
rhoCA
```

```
## [1] 0.7406312
```

O coeficiente de correlação intraclasses (CCI) para a divisão A é aproximadamente 1, 1, 1, 1, 2, 2, 2, 2. O valor mínimo para o CCI é $-\frac{1}{N-1} = -0.333$, o que indica grande ineficiência da AC1S usando a

Avaliação Parcial 2

Questão 1 - solução

b) Para a divisão B temos

```
## divisao B
Ni <- mean(table(CB))           # tamanho de cada conglomerado,
## cria os pares
YlinhaCB1 = permutations( length( Y[CB == 1]), 2, Y[CB == 1] ,set=F)
YlinhaCB2 = permutations( length( Y[CB == 2]), 2, Y[CB == 2] ,set=F)
YlinhaCB = rbind( YlinhaCB1, YlinhaCB2)   # pares de toda a pop.
rhoCB = cor( YlinhaCB[,1], YlinhaCB[,2])  # coef. corr. intraclass
rhoCB
```

```
## [1] -0.2534517
```

O coeficiente de correlação intraclasse (CCI) para a divisão B é aproximadamente -0.2534517 , o que indica uma grande eficiência da $AC1S$ usando a conglomeração A em relação a **AAS**, pois ρ_{int} está próximo do mínimo -0.333 .

c) Não recomendaria na conglomeração A e recomendaria na B . Percebemos

Avaliação Parcial 2

Questão 2 - solução

Do enunciado temos:

- $N = \overline{N} \times M = 4 \times 175 = 700$ pneus, unidades elementares, na população,
- $M = 175$ táxis, conglomerados.
- $m = 10$ táxis foram selecionados com reposição
- T_i número de pneus (dentre os 4 pneus em uso) fora de condições de segurança.

Variável de interesse Y_{ij} : indicadora do pneu j do táxi i estar em condições de segurança.

Avaliação Parcial 2

Questão 2 - solução

```
## exercicio 2
## dados do problema
M <- 175          # no. conglomerados pop.
Ni <- rep(4, 175) # tamanho conglomerados
N <- mean(Ni) * M    # tamanho pop
Ti_amostra <- c(2,4,0,1,2,0,4,1,3,1) # totais obtidos
m <- length(Ti_amostra) # no. conglomerados amostra
Ni_amostra <- Ni[1:m]    # tamanho dos cong.
n <- mean(Ni_amostra) * m # tamanho amostra
```

Avaliação Parcial 2

Questão 2 - solução

```
## a)
PchapeuHT <- (M/m) * sum(Ti_amostra) / N # estimador HT
PchapeuR <- sum(Ti_amostra) / n          # estimador R
```

a) Para estimar pontualmente a proporção de pneus da frota fora de condições, temos dois possíveis estimadores. Aqui conhecemos o tamanho da população N então ambos os estimadores são possíveis. Além disso, lembre que ambos os estimadores são iguais no caso de $N_i = \bar{N}$, temos

$$\hat{P}_{AC1S}^{HT} = \frac{\hat{T}_{AC1S/HT}}{N} = \frac{M}{N} \frac{1}{m} \sum_{i \in a} T_i = \frac{\frac{175}{10} 18}{700} = 0.45$$

ou

$$\hat{P}_{AC1S}^R = \frac{\hat{T}_{AC1S}^R}{N} = \frac{1}{n} \sum_{i \in a} T_i = \frac{\frac{700}{40} 18}{700} = 0.45.$$

Avaliação Parcial 2

Questão 2 - solução

Um intervalo de confiança para \hat{P} é dados por

$$IC_{AC1S}(P; 1 - \alpha) = [\hat{P}_{AC1S} \pm z_{\alpha/2} \sqrt{\hat{Var}_{AC1S}(\hat{P}_{AC1S})}]$$

Assim, temos o erro absoluto, $e = z_{0,05} * \sqrt{\hat{Var}_{AC1S}(\hat{P}_{AC1S})}$

- O estimador não viciado da variância de \hat{P}^{HT} na **AC1S** é dada por:
 - **COM** reposição, $\hat{Var}_{AC1S_c}(\hat{P}^{HT}) = \frac{1}{N^2} \left(1 - \frac{1}{M}\right) \frac{s_{ec}^2}{m} \approx \frac{1}{N^2} \frac{s_{ec}^2}{m}$
 - **SEM** reposição, $\hat{Var}_{AC1S_s}(\hat{P}^{HT}) = \frac{1}{N^2} \left(1 - \frac{m}{M}\right) \frac{s_{ec}^2}{m}$.

$$\text{em que } s_{ec}^2 = \frac{\sum_{i \in a} (T_i - \bar{y}_C)^2}{m-1}.$$

Avaliação Parcial 2

Questão 2 - solução

```
## IC para a proporcao
pbarraC <- sum(Ti_amostra) / m
s2_ec <- 1/(m-1) * sum((Ti_amostra - pbarraC)^2)      # estimativa var
var_PchapeuHT <- (1 / mean(Ni)^2) * (1 - 1/M) * s2_ec / m # estimati
eHT <- -qnorm(0.025) * sqrt(var_PchapeuHT) # erro (absoluto) de est
ICHT <- PchapeuHT + c(-1, 1) * eHT      # intervalo de confianca para
ICHT
```

```
## [1] 0.2219918 0.6780082
```

Então, temos que $IC(P_{AC1S}^{HT}; 0, 95) = [0.2219918; 0.6780082]$. Ou seja, o intervalo de 0.2219918 a 0.6780082 deve conter a proporção de pneus em conformidade com as especificações de segurança de toda a frota de táxis da empresa, com 95% de confiança.

Avaliação Parcial 2

Questão 2 - solução EXTRA

```
## IC para a proporcao - de razao
nbarra <- mean(Ni_amostra)      # estimativa do tamanho da amostra (nu
var_PchapeuR <- (1 / (m * nbarra^2)) * sum((Ti_amostra - PchapeuR * m
eR <- -qnorm(0.025) * sqrt(var_PchapeuR) # erro (absoluto) de estima
ICR <- PchapeuR + c(-1, 1) * eR      # intervalo de confianca p
ICR
```

```
## [1] 0.2213375 0.6786625
```

Avaliação Parcial 2

Questão 2 - solução

b) Usando os resultados como um estudo piloto, queremos calcular o número mínimo de táxis necessário para obter uma estimativa da proporção de pneus fora das condições com erro relativo de 2, 5% e 95% de confiança, considerando **AC1S**

AASc de conglomerados

$$CV = \frac{Var_{ecT}}{\bar{Y}_c},$$

$$m = \frac{z_{\alpha/2}^2 CV^2}{e_r^2}.$$

AASs de conglomerados $CV = \frac{S_{ec}^2}{\bar{Y}_c},$

$$m = \frac{M z_{\alpha/2}^2 CV^2}{z_{\alpha/2}^2 CV^2 + (M - 1) e_r^2}.$$

Avaliação Parcial 2

Questão 3 - solução

```
## exercicio 3
## dados do problema
library(SDaA)
data(agpop)
estados <- unique(agpop$state) # estados dos EUA
M <- length( estados)         # no. de estados
Ni <- aggregate( largef92 ~ state, agpop, length) # no. de cidades e
Ti <- aggregate( largef92 ~ state, agpop, sum)    # totais de fazenda
N <- sum(Ni$largef92)          # no. de cidades
```

- Do problema temos:
 - $\backslash(Y=\backslash)``largef92$: número de fazendas com mais de 1.000 hectares em cada cidade dos EUA;
 - $N = 3078$ fazendas registradas no censo;
 - $M = 50$ estados dos EUA.

Nosso interesse é estimar \overline{Y} : número médio de fazendas com mais de 1.000 hectares por cidade dos EUA;

Avaliação Parcial 2

Questão 3 - solução

a) Usando o meu no. cartão a amostra sob **AC1Ss** de $m = 15$ estados é

```
## (a) selecao de estados
m <- 15 # no. conglomerados
set.seed(00119502) # semente aleatoria com meu no. cartao
(estados_amostra <- sample( estados, m )) # estados selecionados

## [1] PA LA NY OR RI SD KY NM OH VA AK TX GA CT VT
## 50 Levels: AK AL AR AZ CA CO CT DE FL GA HI IA ID IL IN KS KY LA MA MD ...
```

b) A partir da amostra, os tamanhos e totais por estado observados

```
## (b) medidas agregadas
(Ni_amostra <- Ni[Ni$state %in% estados_amostra,"largef92"]) # tama

## [1] 5 8 159 120 64 33 62 88 36 67 5 66 254 98 14
```

```
(Ti_amostra <- Ti[Ni$state %in% estados_amostra,"largef92"]) # tota
```

Avaliação Parcial 2

Questão 3 - solução

Para estimação pontual de \bar{Y} temos os estimadores (qual escolher?)

$$\bar{y}_{AC1S/HT} = \frac{\hat{T}_{AC1S/HT}}{N} = \frac{M}{N} \frac{1}{m} \sum_{i \in a} T_i \quad \text{e} \quad \bar{y}_{AC1S}^R = \frac{\hat{T}_{AC1S}^R}{N} = \frac{1}{n} \sum_{i \in a} T_i$$

```
## estimativa
ybarraAC1S_HT <- (M/m) * sum(Ti_amostra) / N      # estimativa por HT
ybarraAC1S_R <- sum(Ti_amostra) / sum(Ni_amostra) # estimativa tipo R
```

Assim, com base no censo agropecuário de 1992 dos EUA, estimamos que o número médio de fazendas com mais de 1.000 hectares por cidade seja aproximadamente 55.08 (ou 47.14) fazendas.

Avaliação Parcial 2

Questão 3 - solução

Para o IC 95% da média da variável ... sabemos que

$$IC_{AC1S}(\bar{Y}; 1 - \alpha) = [\bar{y}_{AC1S} \mp z_{\alpha/2} \sqrt{\hat{Var}_{AC1S}(\bar{y}_{AC1S})}]$$

em que $\hat{Var}_{AC1S}(\bar{y}_{AC1S}) = \frac{s_{ec}^2}{m}$.

```
## IC
alpha <- 0.05
z_alpha2 <- qnorm(alpha/2)
# varybarraAC1S_HT <-
# erroIC <- z_alpha2 * sqrt() # estimativa por HT
# ICHT <- ybarraAC1S_HT + c(1,-1) * erroIC
# ybarraAC1S_R <- sum(Ti_amostra) / sum(Ni_amostra) # estimativa tipo
```

Interprete os resultados.

c) Produza dois gráficos que descrevam a variável `largef92`: um na

Para casa

- Ler o capítulo 3 da apostila da Profa. Vanessa.
- Ler o capítulo 8 do livro 'Amostragem: Teoria e Prática Usando R'.

Próxima aula

- Acompanhar o material no moodle.
- Amostragem Sistemática
 - Introdução.

Muito obrigado!



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)

Resumo da notação