Amostragem por conglomerados em um único estágio (AC): Parte 1

Prof. Caio Azevedo

Introdução

- Já vimos que uma forma de melhorar os resultados inferenciais consiste na divisão da população em (sub)grupos, amostrando-se, de forma apropriada, dentro de cada um deles (e.g., amostragem estratificada).
- Outras vezes tem-se interesse em estudar (sub)grupos de interesse (estimação em pequenos domínios).
- A amostragem por conglomerados em um único estágio (AC) consiste em :
  - Na divisão de uma população em grupos (chamados de conglomerados).
  - Esta divisão é feita segundo alguma(s) característica(s) conhecida(s) na população sob estudo.

Cont.

- A divisão é feita de modo que os elementos dentro de cada conglomerado sejam diferentes entre si (em geral, os conglomerados também são diferentes entre si, embora essa diferença tenda a ser menor do que dentro de cada conglomerado). Ou seja, cada conglomerado deve ser uma representação da população como um todo.
- Sorteia-se um determinado número de conglomerados (segundo algum plano apropriado, por exemplo AAS_c ou AAS_s) e, de cada um desses conglomerados sorteados, observa-se todos os seus elementos.

Cont.

- Motivação: Quando os sistemas de referência não são adequados e/ou custo de atualizá-los é muito elevado, ou ainda quando a logística para identificar as unidades elementares em campo é cara e/ou consome muito tempo.
- Pode ser mais fácil e /ou menos dispendioso selecionar grupos de unidades elementares (conglomerados).
- Exemplos:
  - Amostra de eleitores pode ser obtida pelo sorteio de um número de domicílios.
  - Amostra de trabalhadores pode ser obtida pelo sorteio de um número de empresas.
  - Estudantes podem ser selecionados por uma amostra de escolas ou classes.

Exemplo

- Considere uma população agrupada em 3 conglomerados, como se segue:

$$\mathcal{U} = \{(1), (2, 3, 4), (5, 6)\} = \{C_1, C_2, C_3\}$$

em que $C_1 = \{1\}$, $C_2 = \{2, 3, 4\}$ e $C_3 = \{5, 6\}$

- O plano amostral adotado consiste em sortear dois conglomerados, sem reposição, e entrevistar todos os elementos do conglomerado.
- Espaço amostral em função dos conglomerados:

$S_C(\mathcal{U}) = \{C_1C_2, C_1C_3, C_2C_1, C_2C_3, C_3C_1, C_3C_2\}$, assim

$S(\mathcal{U}) = \{1234, 156, 2341, 23456, 561, 56234\}$,

$S_C(\mathcal{U}) = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4, \boldsymbol{s}_5, \boldsymbol{s}_5\}$.

Cont.

- Note que, nesse caso, o tamanho da amostra também é uma variável aleatória, $n \in \{3, 4, 5\}$.
- Considere o seguinte vetor de dados (populacionais) $\boldsymbol{d} = (12, 7, 9, 14, 8, 10)'$. Assim $\mu = 10$, $s^2 = 6,8$, $\sigma^2 = \frac{34}{6}$.
- Considere a média amostral $\widehat{\mu}$. Assim temos: $\widehat{\mu}(\boldsymbol{s}_1) = 10,5$, $\widehat{\mu}(\boldsymbol{s}_2) = 10$, $\widehat{\mu}(\boldsymbol{s}_3) = 10,5$, $\widehat{\mu}(\boldsymbol{s}_4) = 9,6$, $\widehat{\mu}(\boldsymbol{s}_5) = 10$ e $\widehat{\mu}(\boldsymbol{s}_6) = 9,6$.
- Podemos provar que $\mathcal{E}(\widehat{\mu}) = 10,03$ e $\mathcal{V}(\widehat{\mu}) = 0,14$ (Exercício).
- Considere as três seguintes possíveis divisões de conglomerados:

Cont.

$$\mathcal{U}_A = \{(2,5),(3,6),(1,4)\} \rightarrow \begin{cases} \boldsymbol{d}_1 = (7,8) & \mu_1 = 7,5 & s_1^2 = 0,5, \\ \boldsymbol{d}_2 = (9,10) & \mu_2 = 9,5 & s_2^2 = 0,5, \\ \boldsymbol{d}_3 = (12,14) & \mu_3 = 13,0 & s_3^2 = 2,0, \end{cases}$$

$$\mathcal{U}_B = \{(2,6),(1,5),(3,4)\} \rightarrow \begin{cases} \boldsymbol{d}_1 = (7,10) & \mu_1 = 8,5 & s_1^2 = 4,5, \\ \boldsymbol{d}_2 = (12,8) & \mu_2 = 10,0 & s_2^2 = 8,5, \\ \boldsymbol{d}_3 = (9,14) & \mu_3 = 11,5 & s_3^2 = 12,5, \end{cases}$$

$$\mathcal{U}_C = \{(2,4),(1,5),(3,6)\} \rightarrow \begin{cases} \boldsymbol{d}_1 = (7,14) & \mu_1 = 10,5 & s_1^2 = 24,5, \\ \boldsymbol{d}_2 = (12,8) & \mu_2 = 10,0 & s_2^2 = 8,0, \\ \boldsymbol{d}_3 = (9,10) & \mu_3 = 9,5 & s_3^2 = 0,5, \end{cases}$$

Plano Amostral

- Sorteia-se um único conglomerado segundo AAS e observa-se as duas unidades pertencentes ao mesmo.
- Nesse caso o tamanho da amostra não é uma variável aleatória.
- Podemos calcular as distribuições amostrais de $\widehat{\mu}$, para cada divisão em conglomerados proposta.

Distribuições amostrais de $\widehat{\mu}$

Divisão A	$\mathcal{E}_A(\widehat{\mu}) = 10$	$\mathcal{V}_A(\widehat{\mu}) = \frac{16}{3}$	
$\widehat{\mu}$:	7,5	9,5	13,0
$P(\widehat{\mu})$:	1/3	1/3	1/3

Divisão B	$\mathcal{E}_B(\widehat{\mu}) = 10$	$\mathcal{V}_B(\widehat{\mu}) = \frac{4,5}{3}$	
$\widehat{\mu}$:	8,5	10,0	11,5
$P(\widehat{\mu})$:	1/3	1/3	1/3

Divisão C	$\mathcal{E}_C(\widehat{\mu}) = 10$	$\mathcal{V}_C(\widehat{\mu}) = \frac{0,5}{3}$	
$\widehat{\mu}$:	9,5	10,0	10,5
$P(\widehat{\mu})$:	1/3	1/3	1/3

Comentários

- Note que $\widehat{\mu}$ é não viciado sob cada uma das três divisões mas, para a situação C, o estimador apresenta a menor variância.
- Neste caso (C), os elementos dentro de cada um dos conglomerados são os mais heterogêneos entre si, o que pode ser medido através da variância média dos conglomerados, notadamente:
  $(A) = (0,5 + 0,5 + 2)/3 = 1$; $(B) = (4,5 + 8 + 12,5)/3 \approx 8,33$;
  $(C) = (24,5 + 8 + 0,5)/3 = 11$.
- Comparando-se amostragem de elementos (AAS) com a de conglomerados (AC), esta última tende a : (i) ter custo de amostragem por elemento menor, (ii) ter maior variância e (iii) maiores problemas para análises estatísticas.

Notações e relações úteis

- Semelhante à estratificação.

$$\begin{aligned}\mathcal{U} &= \{1, 2, ..., N\} \\ &= \{(1,1), ..., (1, B_1), ..., (A, 1), ..., (A, B_A)\} \\ &= \{C_1, C_2, ..., C_A\}\end{aligned}$$

em que

$$C_\alpha = \{(\alpha, 1), ..., (\alpha, i), ..., (\alpha, B_\alpha)\}$$

$$\equiv (\text{conglomerado}, \text{elemento dentro de conglomerado})$$

Diposição dos elementos

Conglomerado	Elementos				
1	y_{11}	$\cdots$	y_{1i}	$\cdots$	y_{1B_1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	
α	$y_{\alpha 1}$	$\cdots$	$y_{\alpha i}$	$\cdots$	$y_{\alpha B_2}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	
A	y_{A1}	$\cdots$	y_{Ai}	$\cdots$	y_{AB_A}

Cont.

- $N = \sum_{\alpha=1}^{A} B_\alpha = A\overline{B}$, $\overline{B} = \frac{N}{A}$, B_α: tamanho do conglomerado α.
- $\tau_\alpha = \sum_{i=1}^{B_\alpha} y_{\alpha i}$ (total populacional do conglomerado α),
- $\tau = \sum_{\alpha=1}^{A} \tau_\alpha = \sum_{\alpha=1}^{A} \sum_{i=1}^{B_\alpha} y_{\alpha i} = A\overline{\tau}$, $\overline{\tau} = \frac{\tau}{A} = \frac{1}{A} \sum_{\alpha=1}^{A} \tau_\alpha$ (total populacional).

Cont.

- $\mu_\alpha = \dfrac{\tau_\alpha}{B_\alpha} = \dfrac{1}{B_\alpha}\sum_{i=1}^{B_\alpha} y_{\alpha i}$ (média populacional do conglomerado α),
- $\mu = \dfrac{\tau}{N} = \dfrac{1}{N}\sum_{\alpha}^{A}\sum_{i=1}^{B_\alpha} y_{\alpha i} = \dfrac{1}{A\overline{B}}\sum_{\alpha=1}^{A}\tau_\alpha = \dfrac{1}{A}\sum_{\alpha}^{A}\dfrac{B_\alpha}{\overline{B}}\mu_\alpha = \dfrac{\overline{\tau}}{\overline{B}}$ (média populacional).
- $\overline{\mu} = \dfrac{1}{A}\sum_{\alpha}^{A}\mu_\alpha$(média das médias dos conglomerados).
- Note que

  $(\mu - \overline{\mu}) = \dfrac{1}{A}\sum_{\alpha=1}^{A}\dfrac{B_\alpha}{\overline{B}}\mu_\alpha - \dfrac{1}{A}\sum_{\alpha}^{A}\mu_\alpha = \dfrac{1}{A}\sum_{\alpha}^{A}\left(\dfrac{B_\alpha}{\overline{B}} - 1\right)\mu_\alpha$ (ou seja, nem sempre μ é igual à $\overline{\mu}$.

Cont.

- $\sigma_\alpha^2 = \frac{1}{B_\alpha} \sum_{i=1}^{B_\alpha} (y_{\alpha i} - \mu_\alpha)^2$ (variância do conglomerado α),
- $\sigma^2 = \frac{1}{N} \sum_{\alpha=1}^{A} \sum_{i=1}^{B_\alpha} (y_{\alpha i} - \mu)^2 =$

  $\frac{1}{N} \sum_{\alpha=1}^{A} \sum_{i=1}^{B_\alpha} (y_{\alpha i} - \mu_\alpha)^2 + \frac{1}{N} \sum_{a=1}^{A} B_\alpha (\mu_\alpha - \mu)^2$ (variância populacional)

  ou seja

  $\sigma^2 =$ variância dentro dos conglomerados $+$

  variância entre os conglomerados $= \sigma_{dc}^2 + \sigma_{ec}^2$,

  em que (próximo slide)

Cont.

- $\sigma^2_{dc} = \frac{1}{N}\sum_{\alpha=1}^{A}\sum_{i=1}^{B_\alpha}(y_{\alpha i} - \mu_\alpha)^2 = \frac{1}{A\overline{B}}\sum_{\alpha=1}^{A}\frac{B_\alpha}{B_\alpha}\sum_{i=1}^{B_\alpha}(y_{\alpha i} - \mu_\alpha)^2 = \frac{1}{A}\sum_{\alpha=1}^{A}\frac{B_\alpha}{\overline{B}}\sigma^2_\alpha$
- $\sigma^2_{ec} = \frac{1}{N}\sum_{\alpha=1}^{A}B_\alpha(\mu_\alpha - \mu)^2 = \frac{1}{A}\sum_{\alpha=1}^{A}\frac{B_\alpha}{\overline{B}}(\mu_\alpha - \mu)^2$
- $\sigma^2_{ec}[\tau] = \frac{1}{A}\sum_{\alpha=1}^{A}(\tau_\alpha - \overline{\tau})^2 = \frac{1}{A}\sum_{\alpha=1}^{A}(B_\alpha\mu_\alpha - \overline{B}\mu)^2 = \frac{\overline{B}^2}{A}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\mu_\alpha - \mu\right)^2 = \overline{B}^2\sigma^2_{ect}$

Cont.

- $\sigma^2_{ect} = \frac{1}{A}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\mu_\alpha - \mu\right)^2$
- $\sigma^2_{eq} = \frac{1}{A}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\right)^2(\mu_\alpha - \mu)^2$
- $\sigma^2_{em} = \frac{1}{A}\sum_{\alpha=1}^{A}(\mu_\alpha - \overline{\mu})^2$
- Sob AAS_s, se necessário, utilizaremos as variâncias populacionais $s^2_{(.)}$, com mudanças adequadas nos respectivos denominadores (como feito antetiormente).

Cont.

- Somas de quadrados

$$SQ[T] = \sum_{\alpha=1}^{A} \sum_{i=1}^{B_\alpha} (y_{\alpha i} - \mu)^2 = N\sigma^2 = A\overline{B}\sigma^2$$

$$SQ[D] = \sum_{\alpha=1}^{A} \sum_{i=1}^{B_\alpha} (y_{\alpha i} - \mu_\alpha)^2 = \sum_{\alpha=1}^{A} B_\alpha \sigma_\alpha^2 = A\overline{B}\sigma_{dc}^2$$

$$SQ[E] = \sum_{\alpha=1}^{A} B_\alpha (\mu_\alpha - \mu)^2 = A\overline{B}\sigma_{ec}^2$$

em que

SQ[T]: soma de quadrados total entre os elementos, SQ[D]: soma de quadrados dentro dos conglomerados, SQ[E]: soma de quadrados entre os elementos. Note que $SQ[T] = SQ[D] + SQ[E]$.

Cont.

- Quando todos os conglomerados tiverem o mesmo tamanho, isto é $B_1 = B_2 = ... = B_A = \overline{B} = B$, teremos que $\frac{B_\alpha}{\overline{B}} = 1$, $\mu = \overline{\mu}$ e

$$\sigma^2_{ec} = \sigma^2_{ect} = \sigma^2_{eq} = \sigma^2_{em} = \frac{1}{A}\sum_{\alpha=1}^{A}(\mu_\alpha - \mu)^2$$

Plano amostral

- Serão sorteados $a < A$ conglomerados, através de um processo AAS_c (exercício: repetir os desenvolvimentos, aqui apresentados, sob AAS_s).
- De cada conglomerado serão analisados todas as unidades populacionais.
- Equivale ao procedimento AAS_c, anteriormente estudado, em que $\mathcal{U}_C = \{C_1, C_2, ..., C_\alpha, ..., C_A\}$.

Cont.

- Quantidades populacionais

$$\boldsymbol{d} = \begin{pmatrix} B_1 & B_2 & \dots & B_\alpha & \dots & B_A \\ \tau_1 & \tau_2 & \dots & \tau_\alpha & \dots & \tau_A \\ \mu_1 & \mu_2 & \dots & \mu_\alpha & \dots & \mu_A \end{pmatrix}$$

- Quantidades amostrais

$$\boldsymbol{D} = \begin{pmatrix} b_1 & b_2 & \dots & b_\alpha & \dots & b_a \\ \widehat{\tau}_1 & \widehat{\tau}_2 & \dots & \widehat{\tau}_\alpha & \dots & \widehat{\tau}_a \\ \widehat{\mu}_1 & \widehat{\mu}_2 & \dots & \widehat{\mu}_\alpha & \dots & \widehat{\mu}_a \end{pmatrix}$$

- Assim, todas as propriedades e resultados derivadas para *AAS* (com reposição e sem reposição) são validas aqui, considerando $n = \sum_{\alpha=1}^{a} b_\alpha$.

Estimadores para a média populacional

- O parâmetro a ser estimado é $\mu = \frac{\tau}{N} = \frac{\overline{\tau}}{\overline{B}} = \frac{\frac{1}{A}\sum_{\alpha=1}^{A}\tau_\alpha}{\frac{1}{A}\sum_{\alpha=1}^{A}B_\alpha}$
- Estimador 1: supõe conhecido o número total N de unidades na população.

$$\widehat{\mu}_{C_1} = \frac{A\widehat{\overline{\tau}}}{A\overline{B}} = \frac{\widehat{\overline{\tau}}}{\overline{B}}, \widehat{\overline{\tau}} = \frac{1}{a}\sum_{\alpha=1}^{a}\widehat{\tau}_\alpha, \overline{B} = \frac{N}{A} = \frac{\sum_{\alpha=1}^{A}B_\alpha}{A}.$$

Estimadores para a média populacional

- Estimador 2: mais indicado quando o total N é desconhecido.

$$\widehat{\mu}_{C_2} = \frac{A\widehat{\overline{\tau}}}{A\widehat{\overline{B}}} = \frac{\widehat{\overline{\tau}}}{\widehat{\overline{B}}}, \widehat{\overline{\tau}} = \frac{1}{a}\sum_{\alpha=1}^{a}\widehat{\tau}_\alpha, \widehat{\overline{B}} = \frac{1}{a}\sum_{\alpha=1}^{a} b_\alpha.$$

- Estimador 3: ignora o fato dos conglomerados terem tamanhos diferentes

$$\widehat{\mu}_{C_3} = \frac{1}{a}\sum_{\alpha=1}^{a}\widehat{\mu}_\alpha.$$

Cont.

- Resultado: Sob AASc (suprimindo o sub índice referente ao plano amostral), temos que
  - $\mathcal{E}(\widehat{\mu}_{C_1}) = \mu, \quad \mathcal{E}(\widehat{\mu}_{C_2}) = \mu + B(\widehat{\mu}_{C_2}), \quad \mathcal{E}(\widehat{\mu}_{C_3}) = \mu + (\overline{\mu} - \mu)$

    em que $B(\widehat{\mu}_2)$ denota o vício do estimador $\widehat{\mu}_2$. Al[em disso:

$$\mathcal{V}(\widehat{\mu}_{C_1}) = \frac{\sigma^2_{ect}}{a} = \frac{1}{aA}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\mu_\alpha - \mu\right)^2, \tag{1}$$

$$EQM(\widehat{\mu}_{C_2}) \approx \mathcal{V}(\widehat{\mu}_2) = \frac{\sigma^2_{eq}}{a} = \frac{1}{aA}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\right)^2(\mu_\alpha - \mu)^2, \tag{2}$$

$$EQM(\widehat{\mu}_{C_3}) = \frac{\sigma^2_{em}}{a} + (\overline{\mu} - \mu)^2 = \frac{1}{aA}\sum_{\alpha=1}^{A}(\mu_\alpha - \overline{\mu})^2 + (\overline{\mu} - \mu)^2 \tag{3}$$

Esboço de demonstrações

- O estimador 1 é função da média artimética simples obtida a partir de uma AASc, dos valores $\tau_1, \tau_2, ..., \tau_A$. O resultado segue.
- Estimador 2: basta lembrar quem são $\boldsymbol{d}$ e $\boldsymbol{D}$ e observar que $\widehat{\mu}_{C_2}$ é um estimador razão.
- Estimador 3: o mesmo raciocínio usado para o estimador 1, sendo que os valores são $\mu_1, \mu_2, ..., \mu_A$.

Estimadores para as variâncias dos estimadores

- $\widehat{\mathcal{V}}(\widehat{\mu}_{C_1}) = \dfrac{1}{a(a-1)} \displaystyle\sum_{\alpha=1}^{a} \left(\frac{B_\alpha}{\overline{B}} \widehat{\mu}_\alpha - \widehat{\mu}_{C_1} \right)^2$
- $\widehat{\mathcal{V}}(\widehat{\mu}_{C_2}) = \dfrac{1}{a(a-1)} \displaystyle\sum_{\alpha=1}^{a} \left(\frac{b_\alpha}{\overline{b}} \right)^2 (\widehat{\mu}_\alpha - \widehat{\mu}_{C_2})^2$, em que $\overline{b} = \dfrac{1}{\alpha} \displaystyle\sum_{\alpha=1}^{a} b_\alpha$.
- $\widehat{\mathcal{V}}(\widehat{\mu}_{C_3}) = \dfrac{1}{a(a-1)} \displaystyle\sum_{\alpha=1}^{a} (\widehat{\mu}_\alpha - \widehat{\mu}_{C_3})^2$
- Sob AAS_c, o primeiro e o terceiro estimadores são não viciados. A prova è semelhante ao resultado anterior.

Cont.

- Nenhum dos 3 estimadores $\widehat{\mu}_{C_1}, \widehat{\mu}_{C_2}, \widehat{\mu}_{C_3}$ tem EQM menor do que os outros dois, sob toda e qualquer circunstância (a não ser, em casos específicos).
- Jessen (1978) afirma que, se o coeficiente de regressão de $\mu_\alpha(\widehat{\mu}_\alpha)$ em função de $B_\alpha(b_\alpha)$, for negativo, positivo ou nulo, deve-se preferir $\widehat{\mu}_{C_1}, \widehat{\mu}_{C_2}$ ou $\widehat{\mu}_{C_3}$, respectivamente.
- Ou seja, ajusta-se o modelo (de regressão)

$$\widehat{\mu}_\alpha = \gamma_0 + \gamma_1 B_\alpha + \epsilon \tag{4}$$

e avalia-se o valor (significância) de $\widehat{\gamma}_1(\gamma_1)$

Conglomerados de igual tamanho

- Quando todos os conglomerados têm o mesmo tamanho B, os três estimadores são iguais à : $\widehat{\mu}_C = \frac{1}{aB}\sum_{\alpha=1}^{a}\sum_{i=1}^{B} y_{\alpha i} = \frac{1}{a}\sum_{\alpha=1}^{a}\widehat{\mu}_\alpha$ com

$$\mathcal{V}(\widehat{\mu}_C) = \frac{\sigma^2_{ec}}{a} = \frac{1}{aA}\sum_{\alpha=1}^{A}(\mu_\alpha - \mu)^2.$$

- Um estimador não viciado para a $\mathcal{V}(\widehat{\mu}_C)$ é dado por

$$\widehat{\mathcal{V}}(\widehat{\mu}_C) = \frac{\widehat{\sigma}^2_{ec}}{a} = \frac{1}{a(a-1)}\sum_{\alpha=1}^{a}(\widehat{\mu}_\alpha - \widehat{\mu}_C)^2.$$

- É importante notar, quando todos os conglomerados têm igual tamanho, que

$$\widehat{\sigma}^2_{ec} = \widehat{\sigma}^2_{ect} = \widehat{\sigma}^2_{eq} = \widehat{\sigma}^2_{em} = \frac{1}{a-1}\sum_{\alpha=1}^{a}(\widehat{\mu}_\alpha - \widehat{\mu}_C)^2$$

Cont.

- Além disso, o estimador $\widehat{\sigma}^2_{dc} = \dfrac{1}{a}\sum_{\alpha=1}^{a} \dfrac{B_\alpha}{\overline{B}}\widehat{\sigma}^2_\alpha$ é não viciado para σ^2_{dc}.
- Quando $\overline{B}$ for desconhecido, substitui-mo-no por $\overline{b}$, o que leva o estimador anterior a ser viciado.
- Se os tamanhos dos conglomerados não variarem muito entre si, então o viés passar a ser pequeno.

Coeficiente de correlação intraclasse

- A eficiência (condução à inferências mais precisas) do processo de divisão (de uma ou mais populações) em conglomerados depende do grau de similaridade de seus elementos.
- É importante criar medidas que indiquem o grau de similaridade dos elementos dentro dos conglomerados.
- Existem várias propostas na literatura, principalmente quando os conglomerados têm tamanhos distintos.
- Usaremos o coeficiente de correlação intraclasse ρ_{int} (link 1, link 2, link 3, link 4).

Processo (algoritmo) para o cálculo do ρ_{int}

- Considere a população dividida em A conglomerados como definido anteriormente.
- Em seguida, forma-se todos os pares de unidades distintas possíveis dentre de cada conglomerado. Por exemplo, para o $\alpha-$ésimo conglomerado seria possível formar $B_\alpha(B_\alpha - 1)$ pares de valores.
- Desse modo, têm-se no total de conglomerados $\sum_{\alpha=1}^{A} B_\alpha(B_\alpha - 1)$ pares do tipo (y_1', y_2'), em que y_1' indica os possíveis valores da primeira posição do par e y_2', o segundo.
- Calcula-se agora com todos esses $\sum_{\alpha=1}^{A} B_\alpha(B_\alpha - 1)$ pares o coeficiente de correlação de Pearson, ou seja $\rho_{int} = \frac{Cov(y_1', y_2')}{DP(y_1')DP(y_2')}$

Diposição dos elementos

Elemento	$(\alpha, 1)$	$(\alpha, 2)$	$\ldots$	(α, i)	$\ldots$	(α, B_α)
$(\alpha, 1)$	-	$(y_{\alpha 1}, y_{\alpha 2})$	$\ldots$	$(y_{\alpha 1}, y_{\alpha i})$	$\ldots$	$(y_{\alpha 1}, y_{\alpha B_\alpha})$
$(\alpha, 2)$	$(y_{\alpha 2}, y_{\alpha 1})$	-	$\ldots$	$(y_{\alpha 2}, y_{\alpha i})$	$\ldots$	$(y_{\alpha 2}, y_{\alpha B_\alpha})$
$\ldots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
(α, i)	$(y_{\alpha i}, y_{\alpha 1})$	$(y_{\alpha i}, y_{\alpha 2})$	$\ldots$	-	$\ldots$	$(y_{\alpha 1}, y_{\alpha B_\alpha})$
$\ldots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
(α, B_α)	$(y_{\alpha B_\alpha}, y_{\alpha 1})$	$(y_{\alpha B_\alpha}, y_{\alpha 2})$	$(y_{\alpha B_\alpha}, y_{\alpha i})$	$\ldots$	-	

(voltando ao) Exemplo

- Divisão A

y_1' :	7	8	9	10	12	14
y_2' :	8	7	10	9	14	12

$\rho_{int} \approx 0,82$

- Divisão B

y_1' :	7	10	12	8	9	14
y_2' :	10	7	8	12	14	9

$\rho_{int} \approx -0,47$

- Divisão C

y_1' :	7	14	12	8	9	10
y_2' :	14	7	8	12	10	9

$\rho_{int} \approx -0,94$

Conglomerados de igual tamanho

- Quando todos os conglomerados têm o mesmo tamanho, vem que
  - $Cov(y_1', y_2') = \frac{1}{AB(B-1)} \sum_{\alpha=1}^{A} \sum_{i \neq j} (y_{\alpha i} - \mu)(y_{\alpha j} - \mu)$
  - $Var(y_1') = var(y_2') = \sigma^2$

$$\rho_{int} = \frac{\sigma_{ec}^2 - \frac{\sigma_{dc}^2}{B-1}}{\sigma^2} \tag{5}$$

Interpretação

- Suponha o caso em que $\sigma^2_\alpha = 0, \forall \alpha$ (máxima homogeneidade dentro dos conglomerados, ou seja, todos os elementos são iguais entre si). Temos que: $\sigma^2_{dc} = 0$ e $\sigma^2 = \sigma^2_{ec}$. Assim $\rho_{int} = 1$, que corresponde ao maior valor possível para ρ_{int}.
- Suponha que agora cada conglomerado seja uma microrepresentação da população, ou seja, $\sigma^2_\alpha = \sigma^2 \rightarrow \sigma^2_{dc} = \sigma^2$, logo $\sigma^2_{ec} = 0$. Assim $\rho_{int} = -\frac{1}{B-1}$
- Ou seja, em geral, quanto mais próximo de -1 for o valor (estimativa) de ρ_{int} melhor terá sido o processo de divisão de conglomerados.

EPA

- Para conglomerados de mesmo tamanho, temos que

$$\mathcal{V}(\widehat{\mu}_C) = \{1 + \rho_{int}(B-1)\}\,\frac{\sigma^2}{aB}$$

- $EPA = \dfrac{\mathcal{V}_{AC_1}(\widehat{\mu}_C)}{\mathcal{V}_{A_1}(\widehat{\mu})} = 1 + \rho_{int}(B-1)$
- Em geral (empirismo) $\rho_{int} > 0$.
- Um estimador para ρ_{int} é dado por:

$$\widehat{\rho}_{int} = \frac{\widehat{\sigma}^2_{ec} - \dfrac{\widehat{\sigma}^2_{dc}}{B-1}}{\widehat{\sigma}^2_{ec} + \widehat{\sigma}^2_{dc}}$$

Conglomerados de tamanhos desiguais

- Com o intuito de obter fórmulas operacionais simples, podemos adaptar a fórmula do coeficiente de correlação intraclasse usando algum estimador específico.
- Note que aparecem variâncias entre (σ^2_{ec}) e dentro (σ^2_{dc}) (d)os conglomerados, na fórmula (5).
- Assim, consoante o estimador (para a média) de interesse, podemos substituir σ^2_{ec} por alguma outra variância entre os conglomerados, veja as (variâncias das) expressões (1), (2), (3).

Conglomerados de tamanhos desiguais

- Por exemplo, para $\widehat{\mu}_{C_2}$, temos que $\mathcal{V}(\widehat{\mu}_{C_2}) = \sigma^2_{eq}/a$. Assim, podemos considerar

$$\rho_{C_2} = \frac{\sigma^2_{eq} - \sigma^2_{dc}/(\overline{B} - 1)}{\sigma^2_{eq} + \sigma^2_{dc}}.$$

- Pode-se provar, utilizando-se a fórmula acima, que

$$\mathcal{V}(\widehat{\mu}_{C_2}) = \left\{1 + \rho_{C_2}(\overline{B} - 1)\right\} \frac{\gamma^2}{a\overline{B}}$$

em que $\gamma^2 = \sigma^2_{eq} + \sigma^2_{dc}$.

Conglomerados de tamanhos desiguais

- Além disso,

$$EPA = \left\{1 + \rho_{C_2}(\overline{B} - 1)\right\} \frac{\gamma^2}{\sigma^2}$$

- Se os tamanhos (dos conglomerados) não variarem muito, então $\gamma^2/\sigma^2 \approx 1$ e, portanto

$$EPA \approx 1 + \rho_{C_2}\left(\overline{B} - 1\right)$$

Exemplo

- Considere a população definida no começo dos slides

$$\mathcal{U} = \{(1), (2, 3, 4), (5, 6)\} = \{C_1, C_2, C_3\}$$

em que $C_1 = \{1\}$, $C_2 = \{2, 3, 4\}$ e $C_3 = \{5, 6\}$,

$\boldsymbol{d} = ((12), (7, 9, 14), (8, 10))$

- Temos que $\mu = 10$, $\sigma^2 = 17/3$, $\overline{\mu} = 31/3$, $\overline{B} = 2$.
- C1: $\mu_1 = 12$, $\sigma_1^2 = 0$, $B_1 = 1$.
- C2: $\mu_2 = 10$, $\sigma_2^2 = 26/3$, $B_2 = 3$.
- C3: $\mu_3 = 9$, $\sigma_3^2 = 1$, $B_3 = 2$.

Cont.

- $\sigma^2_{dc} = 14/3$, $\sigma^2_{ec} = 1$, $\sigma^2 = \sigma^2_{dc} + \sigma^2_{ec}$.
- $\sigma^2_{ect} = 14$, $\sigma^2_{eq} = 2/3$, $\sigma^2_{em} = 14/9$.
- Suponha que o plano amostral consista no sorteio de dois conglomerados com reposição.
- Obteremos os resultados através das fórmulas (página 24 destes slides) bem como das distribuições exatas (link). Assim (próximo slide):

Cont.

- **Através das fórmulas, temos**

$$\begin{aligned}
\mathcal{V}(\widehat{\mu}_{C_1}) &= 7, \mathcal{V}(\widehat{\mu}_{C_2}) = 1/3 = 150/450 \approx 0,33, \mathcal{V}(\widehat{\mu}_{C_3}) = 7/9 \approx 0,78 \\
B(\widehat{\mu}_{C_1}) &= 0, B(\widehat{\mu}_{C_2}) = 1/12 \approx 0,08, B(\widehat{\mu}_{C_3}) = 1/3 \approx 0,33 \\
EQM(\widehat{\mu}_{C_1}) &= 7, EQM(\widehat{\mu}_{C_2}) = 49/144 \approx 0,34, EQM(\widehat{\mu}_{C_3}) = 8/9 \approx 0,89
\end{aligned}$$

- **Através das distribuições exatas, temos**

$$\begin{aligned}
\mathcal{V}(\widehat{\mu}_{C_1}) &= 7, \mathcal{V}(\widehat{\mu}_{C_2}) = 283/450 \approx 0,63, \mathcal{V}(\widehat{\mu}_{C_3}) = 7/9 \approx 0,78 \\
B(\widehat{\mu}_{C_1}) &= 0, B(\widehat{\mu}_{C_2}) = 2/15 \approx 0,13, B(\widehat{\mu}_{C_3}) = 1/3 \approx 0,33 \\
EQM(\widehat{\mu}_{C_1}) &= 7, EQM(\widehat{\mu}_{C_2}) = 97/150 \approx 0,65, EQM(\widehat{\mu}_{C_3}) = 8/9 \approx 0,89
\end{aligned}$$

Cont.

- Nesse caso, o melhor estimador é o $\widehat{\mu}_{C_2}$.
- Entretanto da (Equação (4)) temos que: $\widetilde{\gamma}_0 = 12,333; \widetilde{\gamma}_1 = -1,000$ (o que indica uma superioridade do estimador $\widehat{\mu}_{C_1}$). Se fosse levada em consideração a respectiva significância (p=0,5456), teríamos uma indicação de superioridade do estimador $\widehat{\mu}_{C_3}$).

Cont.

- O sinal do coeficiente angular da Equação (4) deve ser considerada como uma ferramenta adicional na escolha do estimador. Outros fatores como tamanho da amostra/população ("a"/"A"), quantidade de elementos ao longo dos conglomerados selecionados, número de conglomerados, variabilidade intra e entre conglomerados, devem ser considerados.
- Na prática podemos comparar as estimativas dos erros-padrão, variâncias e EQM's (eventualmente usando reamostragem).

Cont.

- Coeficiente de correlação intraclasse

y_1' :	7	7	9	9	14	14	8	10
y_2' :	9	14	14	7	7	9	10	8

$\rho_{int} \approx -0,477$

- Usando a definição adaptada, temos que $\gamma^2 = 2/3 + 14/3 = 16/3$,

$$\rho_{C_2} = \frac{\frac{2}{3} - \frac{14/3}{2-1}}{\frac{16}{3}} = -0,75$$

- $\mathcal{V}(\widehat{\mu}_{C_2}) = \{1 + (-0,75)(2-1)\}\,\frac{16/3}{2 \times 2} = 1/3.$
- Note ainda que $\sigma^2 = \frac{17}{3} \approx \frac{16}{3} \approx \gamma^2$.

Resumo

- Estimadores do coeficiente de correlação intraclasse (conglomerados de tamanhos desiguais):
  - $\rho_{C_2} = \dfrac{\widehat{\sigma}^2_{eq} - \dfrac{\widehat{\sigma}^2_{dc}}{B-1}}{\widehat{\gamma}^2}, \widehat{\gamma}^2 = \widehat{\sigma}^2_{eq} + \widehat{\sigma}^2_{dc}$
  - Desenvolver usando $\widehat{\mu}_{C_1}$ e $\widehat{\mu}_{C_3}$.
  - De uma forma geral, podemos utilizar

$$\widehat{\rho}_{int} = \frac{\text{variância entre conglomerados} - \dfrac{\sigma^2_{dc}}{\overline{B}-1}}{\text{variância entre conglomerados} + \sigma^2_{dc}}$$

  - Também é possível estimar a correlação intraclasse (usando a definição original) através da seguinte função do R: link.
  - No caso de conglomerados com tamanhos (aproximadamente) iguais pode-se usar a função clus.rho.

Simulação (comparação de estimadores)

- Dois estudos de simulação:
  - Estudo 1: Simulou-se R = 5.000 populações divididas em conglomerados e calculou-se, para cada uma delas, $\mathcal{V}(.)$, $B(.)$, $EQM(.) = \mathcal{V}(.) + B^2(.)$, $RQEQM(.) = \sqrt{EQM(.)}$ verdadeiros (populacionais), de $\widehat{\mu}_{C_i}), i = 1, 2, 3$, considerando um plano AAS_s.
  - Estudo 2: Simulou-se uma única população dividida em conglomerados e dela selecionou-se R=5.000 amostras (AAS_s), obtendo-se as estimativas para cada um dos três estimadores $(\widehat{\mu}_{C_i}), i = 1, 2, 3)$.

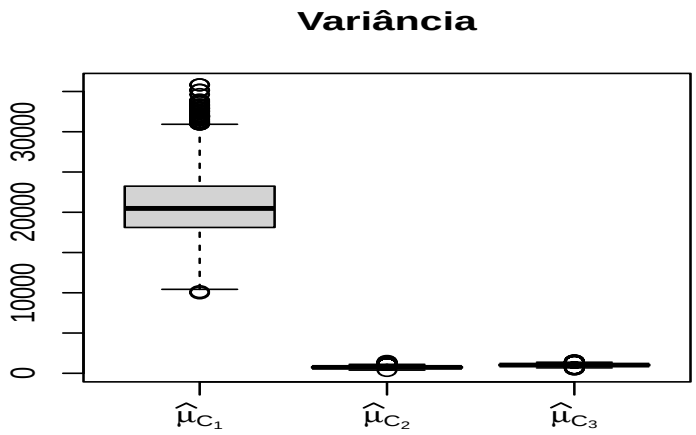
Simulação (comparação de estimadores)

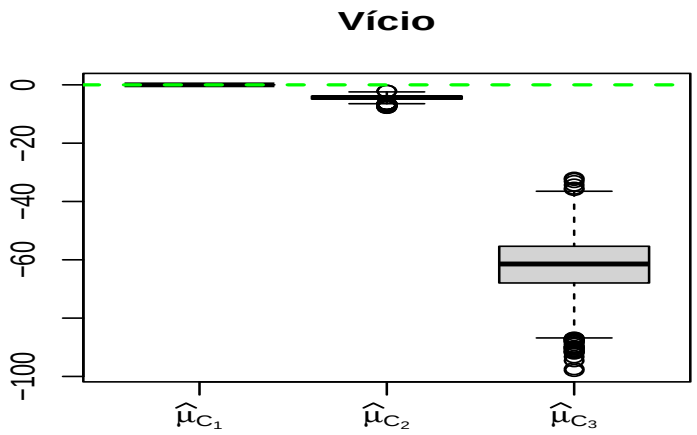
- Estrutura geral (comum):
  - As fórmulas das variâncias populacionais são aquelas apresentadas nestes slides (pag. de 15 a 17), dividindo-se por $A-1$ ao invés de A, ou seja, utilizaremos $s^2_{(.)}$.
  - Também, utilizaremos as fórmulas das variâncias dos estimadores sob AAS_s, ou seja $\mathcal{V}_{A_2}\left(\widehat{\mu}_{C_1}\right) = (1-f)\dfrac{s^2_{ect}}{a}$, $\mathcal{V}_{A_2}\left(\widehat{\mu}_{C_2}\right) = (1-f)\dfrac{s^2_{eq}}{a}$, $\mathcal{V}_{A_2}\left(\widehat{\mu}_{C_3}\right) = (1-f)\dfrac{s^2_{em}}{a}$ $f = a/A$.

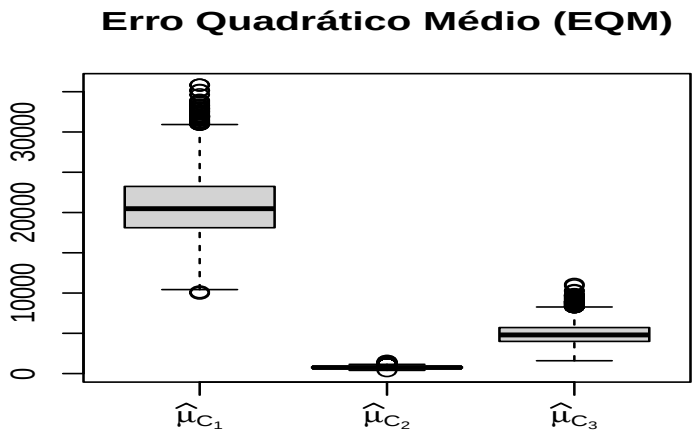
Simulação (comparação de estimadores)

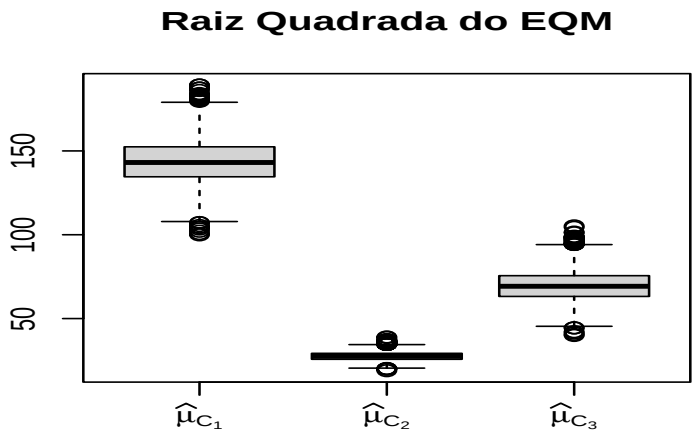
- (cont.) Estrutura geral (comum):
  - Além disso, os "viéses" de $\widehat{\mu}_{C_2}$ (usando a fórmula da página 20 desse link, devidamente adaptada) e de $\widehat{\mu}_{C_3}$ (usando a Equação (3) destes slides).
  - a= 10, A=50, $\sigma^2_\alpha \overset{iid}{\sim} U(200, 400)$, $B_\alpha \overset{iid}{\sim} \lfloor U(10, 500) \rceil$, $\mu_\alpha = 500 + \gamma_1 B_\alpha + \xi_\alpha$, $\xi_\alpha \overset{iid}{\sim} N(0; 0,5)$ e $y_{\alpha i} \overset{ind}{\sim} N(\mu_\alpha, \sigma^2_\alpha)$, $i = 1, 2, ..., B_\alpha$, $\gamma_1 \in \{-0,8; 0; 0,8\}$.

Resultados: estudo 1 - $\gamma_1 = 0,8$









Resulados: estudo 1 - $\gamma_1 = 0,8$ (distribuição da variância)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	9990,77	352,98	541,09
25%	18113,13	646,52	934,95
50%	20476,92	734,20	1020,05
75%	23240,13	833,23	1112,76
100%	35824,45	1473,76	1543,20

Resultados: estudo 1 - $\gamma_1 = 0,8$ (distribuição do vício)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	0,00	-7,73	-97,95
25%	0,00	-4,89	-67,95
50%	0,00	-4,33	-61,44
75%	0,00	-3,84	-55,36
100%	0,00	-2,17	-32,01

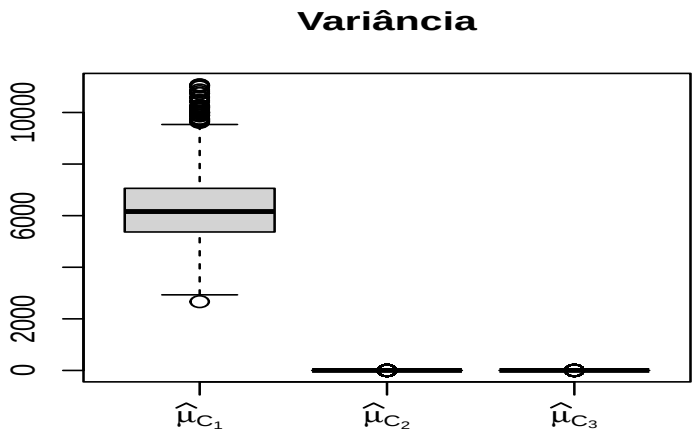
Resultados: estudo 1 - $\gamma_1 = 0,8$ (distribuição do EQM)

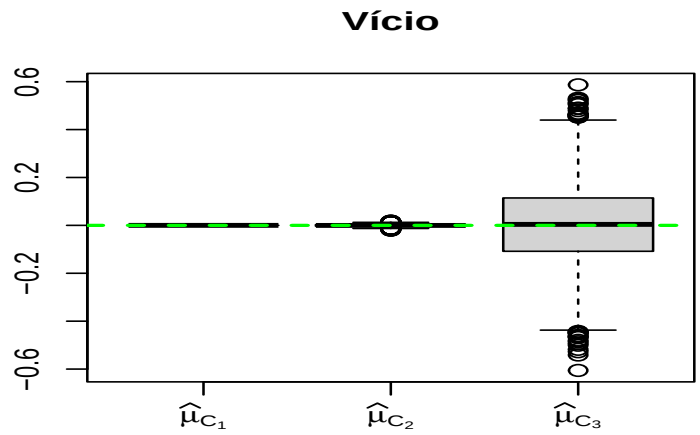
Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	9990,77	357,71	1595,14
25%	18113,13	661,37	3994,78
50%	20476,92	753,13	4798,27
75%	23240,13	856,73	5713,95
100%	35824,45	1523,55	11080,95

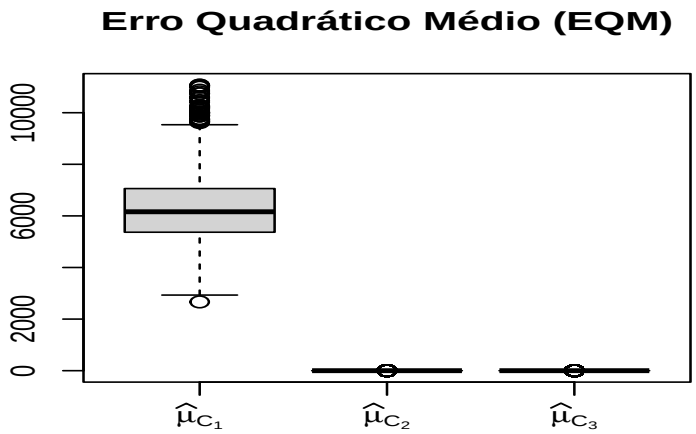
Resultados: estudo 1 - $\gamma_1 = 0,8$ (distribuição do RQEQM)

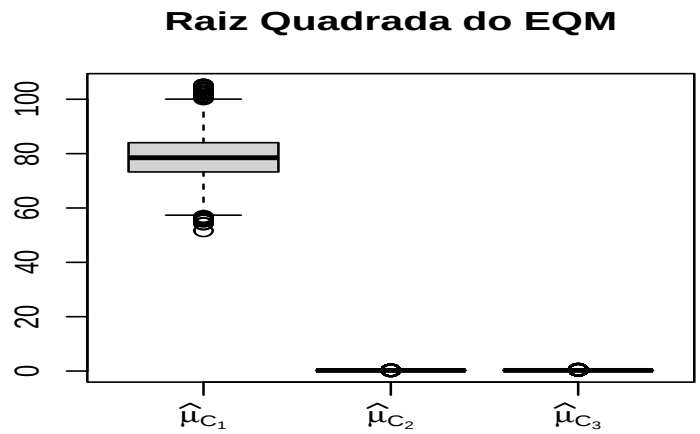
Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	99,95	18,91	39,94
25%	134,59	25,72	63,20
50%	143,10	27,44	69,27
75%	152,45	29,27	75,59
100%	189,27	39,03	105,27

Resultados: estudo 1 - $\gamma_1 = 0,0$









Resulados: estudo 1 - $\gamma_1 = 0,0$ (distribuição da variância)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	2665,94	0,02	0,02
25%	5367,25	0,04	0,03
50%	6160,63	0,05	0,04
75%	7059,29	0,06	0,05
100%	11070,82	0,14	0,07

Resultados: estudo 1 - $\gamma_1 = 0,0$ (distribuição do vício)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	0,00	-0,02	-0,61
25%	0,00	0,00	-0,11
50%	0,00	0,00	0,00
75%	0,00	0,00	0,11
100%	0,00	0,02	0,59

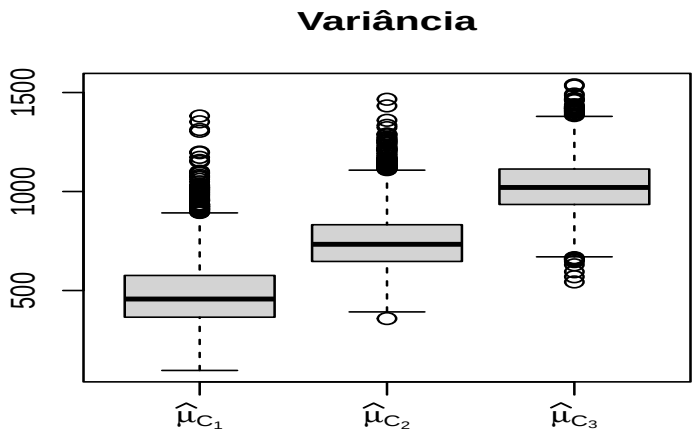
Resultados: estudo 1 - $\gamma_1 = 0,0$ (distribuição do EQM)

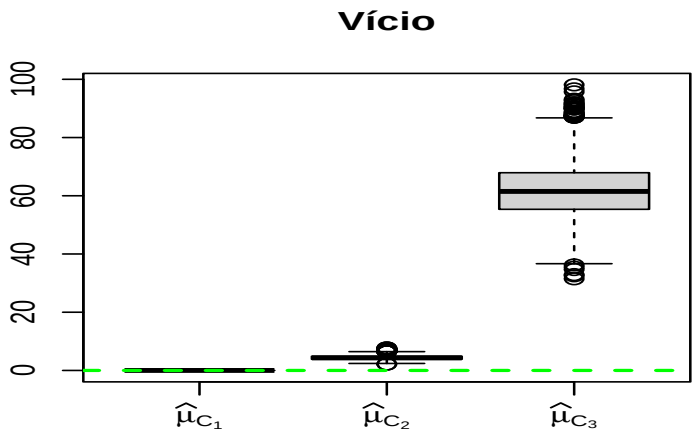
Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	2665,94	0,02	0,02
25%	5367,25	0,04	0,04
50%	6160,63	0,05	0,05
75%	7059,29	0,06	0,08
100%	11070,82	0,14	0,40

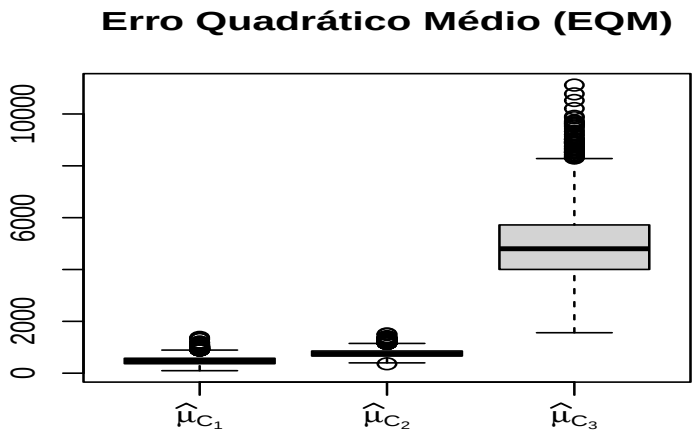
Resultados: estudo 1 - $\gamma_1 = 0,0$ (distribuição do RQEQM)

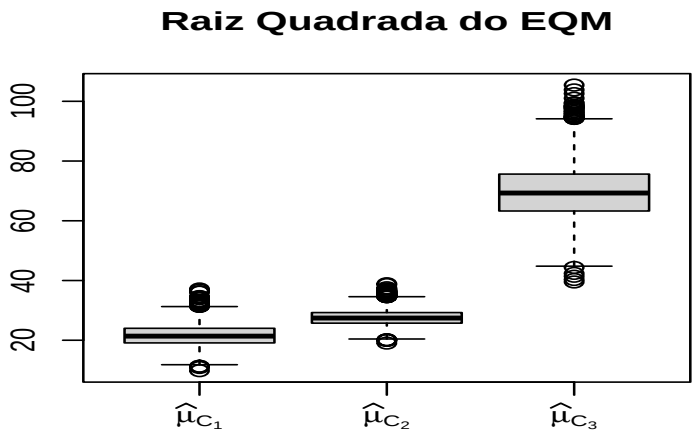
Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	51,63	0,14	0,13
25%	73,26	0,21	0,21
50%	78,49	0,23	0,23
75%	84,02	0,25	0,28
100%	105,22	0,37	0,63

Resultados: estudo 1 - $\gamma_1 = -0,8$









Resulados: estudo 1 - $\gamma_1 = -0,8$ (distribuição da variância)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	96,17	358,20	543,33
25%	365,25	646,75	934,62
50%	457,02	733,41	1020,47
75%	575,90	832,20	1113,52
100%	1381,46	1466,00	1538,78

Resultados: estudo 1 - $\gamma_1 = -0,8$ (distribuição do vício)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	0,00	2,18	31,49
25%	0,00	3,84	55,35
50%	0,00	4,33	61,49
75%	0,00	4,89	67,92
100%	0,00	7,73	98,10

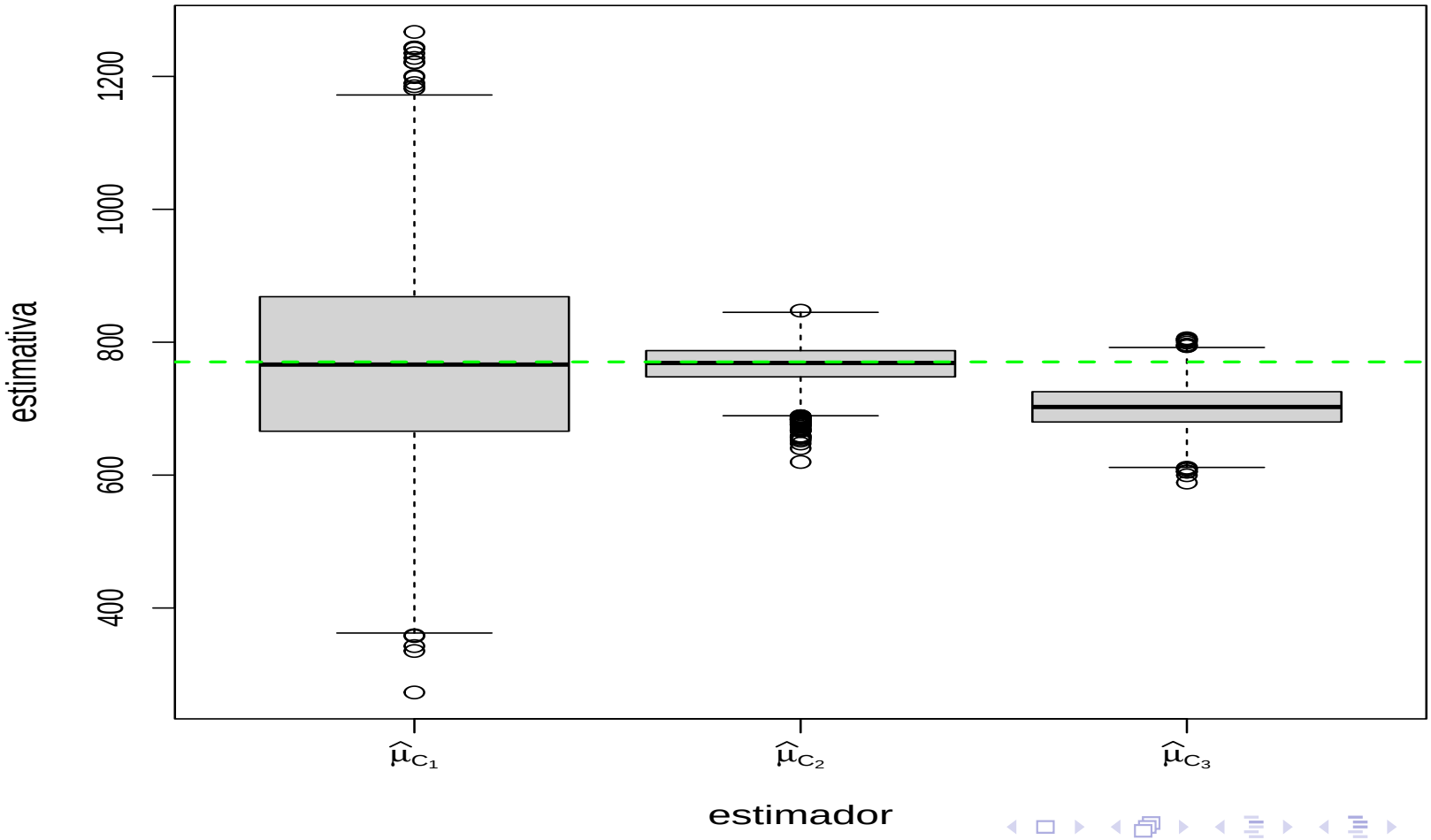
Resultados: estudo 1 - $\gamma_1 = -0,8$ (distribuição do EQM)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	96,17	362,97	1560,96
25%	365,25	661,36	4002,81
50%	457,02	752,64	4800,76
75%	575,90	856,33	5719,56
100%	1381,46	1515,64	11115,12

Resultados: estudo 1 - $\gamma_1 = -0,8$ (distribuição do RQEQM)

Quantil	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
0%	9,81	19,05	39,51
25%	19,11	25,72	63,27
50%	21,38	27,43	69,29
75%	24,00	29,26	75,63
100%	37,17	38,93	105,43

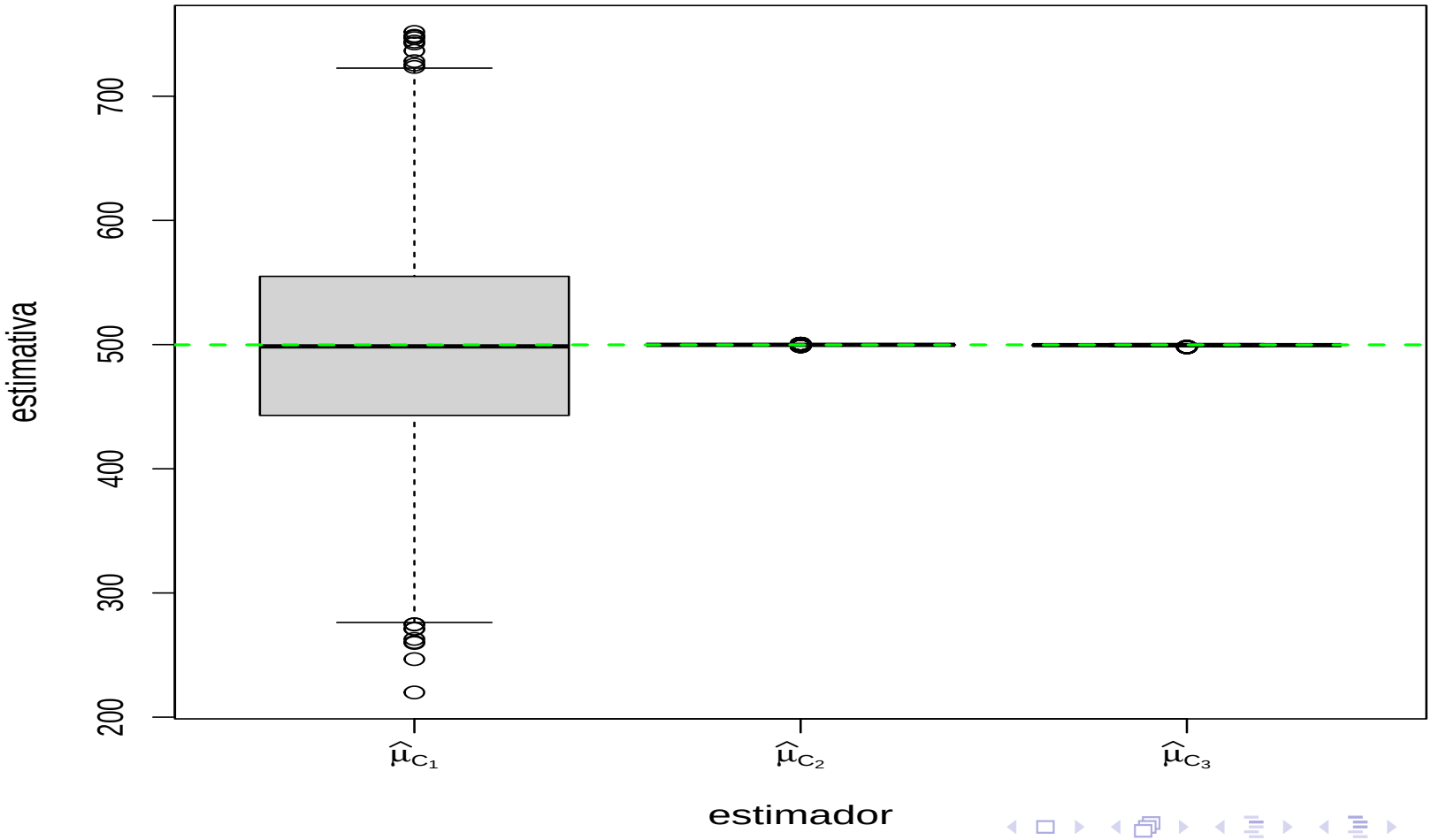
Resultados: estudo 2 - $\gamma_1 = 0,8$



Resulados: estudo 2 - $\gamma_1 = 0,8$ (medidas de acurácia)

Estatística	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
Média	768,72	766,05	702,69
Vício	-1,63	-4,31	-67,67
Variância	21853,59	874,82	1098,81
EQM	21856,26	893,35	5678,28
REQM	147,84	29,89	75,35

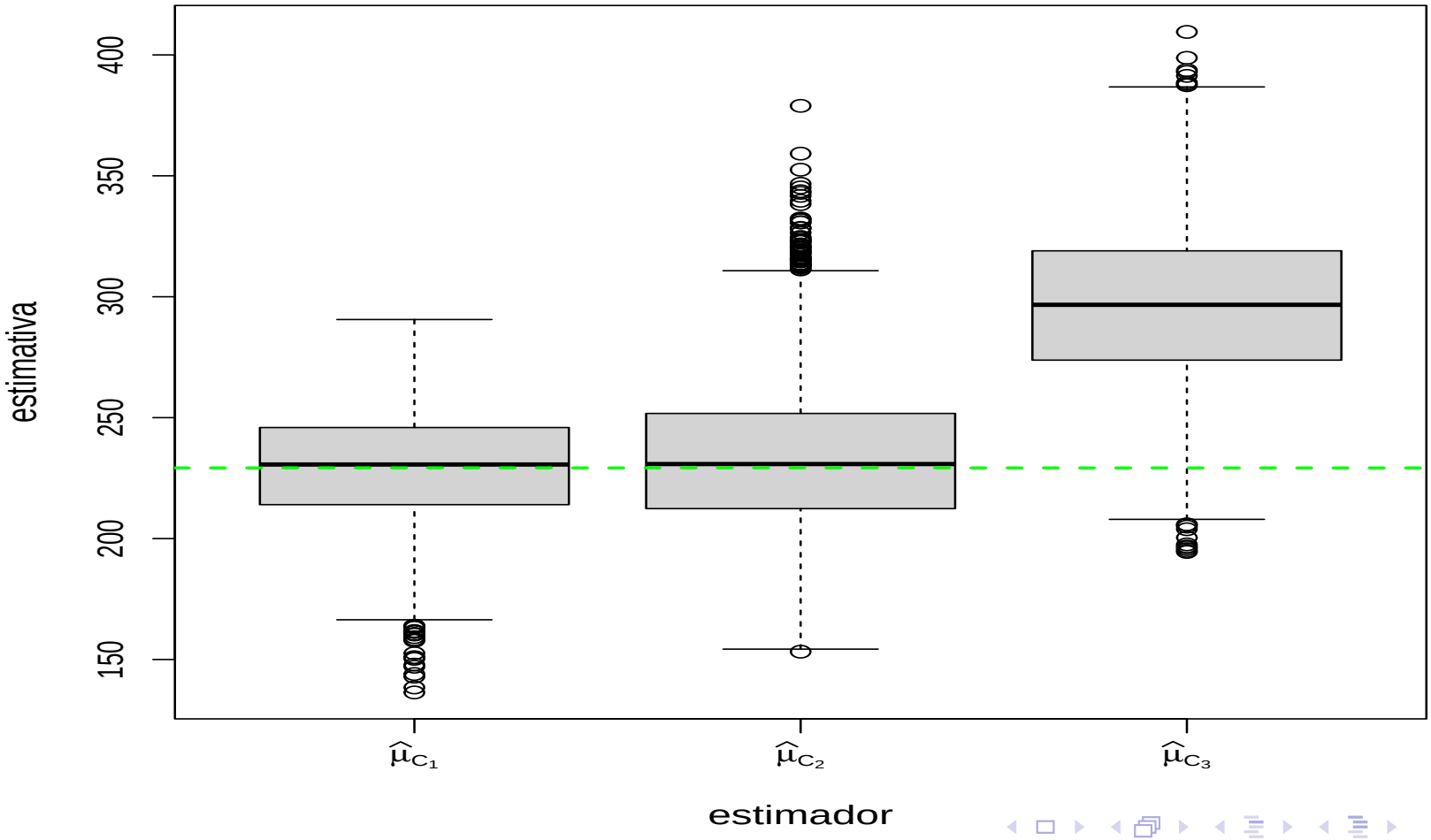
Resultados: estudo 2 - $\gamma_1 = 0,0$



Resulados: estudo 2 - $\gamma_1 = 0,0$ (medidas de acurácia)

Estatística	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
Média	499,10	499,79	499,52
Vício	-0,70	-0,01	-0,28
Variância	6582,06	0,09	0,22
EQM	6582,55	0,09	0,30
REQM	81,13	0,30	0,55

Resultados: estudo 2 - $\gamma_1 = -0,8$



Resulados: estudo 2 - $\gamma_1 = -0,8$ (medidas de acurácia)

Estatística	$\widehat{\mu}_{C_1}$	$\widehat{\mu}_{C_2}$	$\widehat{\mu}_{C_3}$
Média	229,48	233,53	296,35
Vício	0,24	4,29	67,11
Variância	541,93	864,27	1079,06
EQM	541,99	882,65	5582,17
REQM	23,28	29,71	74,71

Comentários

- Com efeito, os estimadores $\widehat{\mu}_{C_1}$ e $\widehat{\mu}_{C_2}$ apresentaram melhor performance quando $\gamma_1 < 0$ e $\gamma_1 > 0$, respectivamente.
- No caso em que $\gamma_1 = 0$ os estimadores $\widehat{\mu}_{C_2}$ e $\widehat{\mu}_{C_3}$ apresentaram desempenho equivalente, com uma leve superioridade para o primeiro. Provavelmente, este resultado ocorreu devido ao fato de que os conglomerados apresentam tamanhos bem diferentes.
- Exercício: realizar simulações considerando outros cenários de interesse como, por exemplo considerando conglomerados de tamanhos parecidos e/ou variâncias dentro e entre conglomerados, menores.