MAT02036 - Amostragem 2

Aula 14 - Amostragem por Conglomerados - Correlação Intraclasse

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 📀

Amostragem por Conglomerados em 1 Estágio Simples

Exercício

Um certo país possui $M = 10$ companhias aéreas com N_i aviões cada. As milhas (*em milhares*) de cada avião (y_{ij}) num determinado período de tempo foram registradas.

Cia (i)	**No. aviões** (N_i)	T_i	$\overline{Y}_i$
1	10	40	4
2	15	75	5
3	15	75	5
4	15	60	4
5	10	60	6
6	15	90	6
7	15	75	5
8	10	70	7
9	10	40	4
10	15	90	6

Aula passada

Amostragem por Conglomerados em 1 Estágio Simples

Exercício

a. Calcule os parâmetros total (T) e média, individual $(\overline{Y})$ e por conglomerados $(\overline{Y}_C)$ e a variância entre totaia dos conglomerados S^2_{ec}.

b*. Calcule o viés dos estimadores **HT** e de **razão** para o total, $\widehat{T}$ (ou média, $\overline{y}$), são não viesados para os respectivos parâmetros que se destinam a estimar, T e $\overline{Y}$. (Obs. mostrar analiticamente ou com os dados do exercício)

b. Assumindo o plano **AC1S** com **AASs** de conglomerados, para amostras de tamanho $m = 4$, calcule a variância do estimador natural (HT) do total, $Var_{AC1S}\left(\widehat{T}^{HT}\right)$, e a variância do estimador da média, $Var_{AC1S}\left(\overline{y}^{HT}\right)$.

c. Repetir o item (b) para estimador o estimador de razão.

d. Escolha um estimador para o total (ou para a média), selecione uma amostra e estime o parâmetro com base na amostra observada.

Efeito do plano amostral

Efeito do plano amostral

- O **Efeito do Plano Amostral - EPA** é uma medida para comparar a **eficiência** de duas estratégias, E_1 e E_2, formadas pelas combinações de **plano amostral** e **estimador**, para um **mesmo tamanho de amostra**.

$$EPA(E_1;\, E_2) = V_{E_1}(\hat{\theta}_1)/V_{E_2}(\hat{\theta}_2)$$

- O termo original em inglês é **Design Effect - deff** e foi sugerido por @Kish1965.

- Outra medida que dá uma indicação **semelhante** ao **EPA** é o **Fator do Plano Amostral - FPA**, que vem do inglês *Design Factor*, definido como:

$$FPA(E_1;\, E_2) = \sqrt{EPA(E_1;\, E_2)} = DP_{E_1}(\hat{\theta}_1)/DP_{E_2}(\hat{\theta}_2)$$

- O FPA compara diretamente o *desvio padrão* dos estimadores sob duas estratégias diferentes de amostragem.
  - É mais comum o uso do EPA que do FPA, sendo o FPA mais diretamente relacionado com a margem de erro das estimativas, enquanto o uso do EPA é mais conveniente quando se trata de planejar e dimensionar amostras.

Efeito do plano amostral

Exemplo - Efeito do plano amostral ao estimar a média populacional por unidade elementar, através do estimador HT com amostragem conglomerada simples em um estágio, em relação ao uso de uma AAS de igual tamanho.

Neste caso, as duas estratégias cuja eficiência se quer comparar são:

- **Estratégia 1**: Amostragem conglomerada em um estágio simples - AC1S, com o estimador natural $\overline{y}_{AC1S/HT}$.
- **Estratégia 2**: Amostragem aleatória simples - AAS de mesmo tamanho total (n), com o estimador usual de média $\overline{y} = \frac{1}{n} \sum_{i \in s} y_i$.

O efeito do plano amostral (neste caso, conglomeração) ao estimar a média populacional por unidade elementar é:

$$EPA(AC1S/HT;\ AAS) = \frac{V_{AC1S}(\overline{y}_{AC1S/HT})}{V_{AAS}(\overline{y})}$$

Efeito do plano amostral

O EPA mede o quanto a variância do estimador é maior (ou menor) por usar, neste caso, AC1S em lugar de AAS.

- $EPA < 1 \Rightarrow$ *ganho de precisão*, devido ao uso de amostragem conglomerada.

- $EPA = 1 \Rightarrow$ *mesma precisão*, não há diferença de precisão, pode-se optar pelo plano operacionalmente mais vantajoso.

- $EPA > 1 \Rightarrow$ *perda de precisão*, devido ao uso de amostragem conglomerada.

Um valor de $EPA = 5$, por exemplo, indicaria que a variância sob amostragem conglomerada seria cinco vezes maior que a variância de uma AAS de igual tamanho total.

Efeito do plano amostral

Exemplo 2 - Efeito do plano amostral ao estimar a média populacional por unidade elementar, através do estimador tipo razão com AC1S em relação ao uso da AAS.

- **Estratégia 1**: Amostragem conglomerada em um estágio simples - AC1S, com estimador tipo razão $\overline{y}_{AC1S}^{R} = \frac{1}{n}\sum_{i \in a} Y_i$ para a média.

- **Estratégia 2**: Amostragem aleatória simples - AAS de mesmo tamanho total (n), com o estimador usual de média $\overline{y} = \frac{1}{n}\sum_{i \in s} y_i$.

O efeito do plano amostral (neste caso, conglomeração) ao estimar a média populacional por unidade elementar é:

$$EPA(AC1S^R;\, AAS) = \frac{V_{AC1S}(\overline{y}_{AC1S}^{R})}{V_{AAS}(\overline{y})}$$

Nota: Os estimadores pontuais são idênticos; somente os planos amostrais (e as variâncias) são diferentes.

Efeito do plano e Correlação Intraclasse

Coeficiente de Correlação Intraclasse

Vimos que a eficiência na **AC1S** depende do grau de similaridade dos seus elementos.

- Uma medida para indicar esse grau de similaridade é o coeficiente de correlação intraclasse (**CCI**), *intraclass correlation coeficient* ou *intracluster correlation coeficient* (**ICC**).

- Considere a população dividida em M conglomerados:
    - Dentro do i-ésimo conglomerado existem $N_i(N_i - 1)$ pares de valores distintos da variável Y.

Elemento	$(i, 1)$	$(i, 1)$	$\dots$	(i, j)	$\dots$	(i, N_i)
$(i, 1)$	-	(Y_{i1}, Y_{i2})	$\dots$	(Y_{i1}, Y_{ij})	$\dots$	(Y_{i1}, Y_{iN_i})
$(i, 2)$	(Y_{i2}, Y_{i1})	-	$\dots$	(Y_{i2}, Y_{ij})	$\dots$	(Y_{i2}, Y_{iN_i})
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(i, j)	(Y_{ij}, Y_{i1})	(Y_{ij}, Y_{i2})	$\dots$	-	$\dots$	(Y_{ij}, Y_{iN_i})
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(i, N_i)	(Y_{iN_i}, Y_{i1})	(Y_{iN_i}, Y_{i2})	$\dots$	(Y_{iN_i}, Y_{ij})	$\dots$	-

Coeficiente de Correlação Intraclasse

- O **CCI** é coeficiente de correlação de Pearson para todos os $\sum_{i \in C} N_i(N_i - 1)$ pares do tipo (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.

Exemplo: (cont. Exemplo slides `Aula 12`)

Considere a população de tamanho $N = 6$ agrupada em $M = 3$ conglomerados, de três maneiras diferentes:

$$\boldsymbol{Y}_A = ((7, 8); (9, 10); (12, 14))$$

$$\boldsymbol{Y}_B = ((7, 10); (12, 8); (9, 14))$$

$$\boldsymbol{Y}_C = ((7, 14); (12, 8); (9, 10))$$

Coeficiente de Correlação Intraclasse

Tamanhos dos conglomerados *iguais*

- Se $N_i = \overline{N},\ \forall\, i = 1, \ldots, M$, sob **AASc** dentro dos conglomerados,

  $$Cov(Y_1', Y_2') = \frac{\sum_{i \in C} \sum_{j \in C_i} \sum_{k \neq j \in C_i} \left(y_{ij} - \overline{Y}\right)\left(y_{ik} - \overline{Y}\right)}{M\overline{N}(\overline{N} - 1)}$$

  e

  $$Var(Y_1') = Var(Y_2') = Var_y,$$

  então

  $$\rho = \frac{Var_{ec} - \frac{Var_{dc}}{\overline{N} - 1}}{Var_y}$$

Mostrar???

Coeficiente de Correlação Intraclasse

Tamanhos dos conglomerados *iguais*

- Sob **AASs** dentro dos conglomerados, ρ_{int} é dado por (@Cochran1977, página 242):

$$\rho_{int} = \frac{2\sum_{j\in C_i}\sum_{k\neq j\in C_i}\left(y_{ij}-\overline{Y}\right)\left(y_{ik}-\overline{Y}\right)}{\left(\overline{N}-1\right)\left(M\overline{N}-1\right)S_y^2} = \frac{(M-1)\overline{N}S_{ec}^2-(M\overline{N}-1)S^2}{(M\overline{N}-1)(\overline{N}-1)S^2}$$

em que $S_{ec}^2 = \frac{1}{M-1}\sum_{i\in C}\left(T_i-\overline{Y}_C\right)^2$ é a variância entre os **totais** dos conglomerados.

- Se o termo $1/M$ é negligenciável $\rho_{int} \doteq \frac{S_{ec}^2-S_y^2}{(\overline{N}-1)S_y^2}$.

Coeficiente de Correlação Intraclasse

Tamanhos dos conglomerados *iguais*

Quais o Valores mínimos e máximos de ρ_{int}? (mostrar❓)

(ver Hansen, Hurvitz and Madow(1953), "measure of homogeneity of the cluster")

1. se $S_i^2 = 0, \forall i$, então $S_{dc}^2 = 0$ e $S^2 = (M-1)S_{ec}^2/(M\overline{N}-1)$ assim $\rho = 1$ (mostrar❓).
2. se $S_{dc}^2 = S^2$, então $S_{ec}^2 = 0$ assim $\rho = -\frac{1}{\overline{N}-1}$ (mostrar ❓).

Coeficiente de Correlação Intraclasse

Tamanhos dos conglomerados *iguais*

- Sob **AASs** de conglomerados, lembrando, variância **total** é dada por:

$$S_y^2 = \frac{1}{N-1} \sum_{i \in C} \sum_{j \in C_i} \left(y_{ij} - \overline{Y} \right)^2$$

- Vimos que a variância **total** também pode ser expressa em função das variâncias *entre* conglomerados, S_{ec}^2, e *dentro* dos conglomerados, S_{dc}^2, através da expressão:

$$S_y^2 = \frac{(\overline{N} - 1) M S_{dc}^2 + \overline{N}(M-1) \overline{S}_{ec}^2}{M\overline{N} - 1}$$

onde, $\overline{S}_{ec}^2 = \frac{S_{ec}^2}{\overline{N}}$.

Coeficiente de Correlação Intraclasse

Tamanhos dos conglomerados *iguais*

- Sob $AASs$ de conglomerados

$$EPA(AC1S^R;\ AAS) \doteq 1 + (\overline{N} - 1)\rho$$

A expressão para o $EPA(AC1S^R;\ AAS)$ resulta do uso das expressões de acordo com @Cochran1977, página 241:

$$Var_{AC1S}(\overline{y}^R_{AC1S}) \doteq \left(\frac{1}{m\overline{N}} - \frac{1}{M\overline{N}}\right) S^2_y[1 + (\overline{N} - 1)\rho]$$

$$Var_{AAS}(\overline{y}) = \left(\frac{1}{m\overline{N}} - \frac{1}{M\overline{N}}\right) S^2_y$$

Coeficiente de Correlação Intraclasse

Algumas considerações relacionadas com a variação do EPA para AC1S:

1. Se os conglomerados tiverem variância dentro grande, isto é, se $S^2_{dc} \doteq S^2_y$, então $\rho \doteq 0$ e portanto, $EPA(AC1S^R;\ AAS) \doteq 1 + (\overline{N} - 1) \times 0 = 1$.

- Nesse caso, não ocorreria perda de precisão devido ao uso de amostragem conglomerada.

- Em muitas aplicações práticas, $\rho > 0$, porque os conglomerados tendem a ser mais homogêneos internamente do que a população em geral. + Consequência: $EPA(AC1S^R;\ AAS) > 1$ na maioria das vezes.

2. Raramente $\rho < 0$, caso em que **AC1S** seria mais eficiente que **AAS**.

3. Num caso extremo, $\rho = 1$ e portanto $EPA(AC1S^R;\ AAS) = \overline{N}$ e $Var_{AC1S}(\overline{y}^R_{AC1S}) = EPA(AC1S^R;\ AAS)\, Var_{AAS}(\overline{y}) \doteq \overline{N}\,\frac{S^2_y}{m\overline{N}} = \frac{S^2_y}{m}$

- Nesse caso, a precisão da amostra conglomerada de tamanho total igual a $m\overline{N}$ é equivalente apenas àquela obtida com uma amostra aleatória simples de tamanho m!!!

Coeficiente de Correlação Intraclasse

Estimação

- Numa **amostra** retirada **com reposição**, o **coeficiente de correlação intraclasse** pode ser **estimado** por:

$$r = \frac{s_{ec}^2 - \frac{s_{dc}^2}{\overline{N}}}{s_{ec}^2 + s_{dc}^2}$$

em que

$$s_{ec}^2 = \frac{1}{m-1} \sum_{i \in a} \left(\overline{Y}_i - \overline{y}^{HT} \right)^2 \quad \text{e} \quad s_{dc}^2 = \frac{1}{m} \sum_{i \in a} \frac{N_i}{\overline{N}} Var_i$$

Coeficiente de Correlação Intraclasse

Exemplo - continuação: Considerando a divisão A:

Para conglomerados de tamanhos diferentes.

Exemplo - Bolfarine e Bussab, apostila pg. 35

Considere a população de tamanho $N = 6$ agrupada em $M = 3$ conglomerados da seguinte forma: $Y = ((12); (7, 9, 14); (8, 10))$

a) Veja neste exemplo que $\frac{\gamma^2}{\sigma^2}$.

b) Calcule o CCI pela definição e pelo método proposto do Bolfarine e Bussab.

Para casa 🏠

- Fazer a lista 2 de exercícios.
- Ler o capítulo 2 da apostila da Profa. Vanessa.
- Rever os slides.

Próxima aula 📊

- Acompanhar o material no moodle.

Amostragem por Conglomerados

- Tamanho de amostra e Intervalos de confiança

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