

Amostragem por conglomerados em dois estágios (AC2E): Um introdução

Prof. Caio Azevedo

Motivação

- Conglomerados homogêneos (elementos dentro de cada conglomerado muito parecidos entre si) o uso da AC (completa) (seleccionando-se todos os elementos de cada conglomerado) se torna menos indicada.
- Dados que os elementos são muito parecidos entre si, eles trarão a mesma informação, essencialmente.

Motivação

- Considere que os elementos dentro de cada conglomerado são todos iguais entre si. Basta selecionar um elemento dentro de cada conglomerado.
- Uma alternativa: selecionar conglomerados e, destes, selecionar um determinado número de elementos.
- Observações: as definições e notações são, essencialmente, as mesmas definidas para o plano AC (http://www.ime.unicamp.br/~cnaber/aula_AC_Amost_2S_2018.pdf, http://www.ime.unicamp.br/~cnaber/aula_AC_P2_Amost_2S_2018.pdf)

Descrição do plano de Amostragem por conglomerados em dois estágios (AC2E)

- Após a população estar agrupada em A conglomerados, procede-se da seguinte forma:
 - Selecionam-se, aleatoriamente, no primeiro estágio, a conglomerados, segundo algum plano amostral.
 - De cada conglomerado selecionado, sorteiam-se b_{α} elementos, segundo o mesmo ou outro plano amostral.
 - Pode-se usar em cada estágio AAS_c ou AAS_s .
- Consideraremos, somente, AAS_c no dois estágios. Veja as referências, para outras combinações. Exercício: repetir os desenvolvimentos considerando AAS_s em cada estágio.

Exemplo

- Considere uma população agrupada em 3 conglomerados, como se segue (o mesmo visto em http://www.ime.unicamp.br/~cnaber/aula_AC_Amost_2S_2018.pdf):

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

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 de cada conglomerado sorteado, sortear-se um elemento com igual probabilidade.
- Espaço amostral: construção em dois estágios.

Exemplo

- Espaço amostral em função dos conglomerados (primeiro estágio):

$$\begin{aligned}S_C(\mathcal{U}) &= \{C_1 C_2, C_1 C_3, C_2 C_1, C_2 C_3, C_3 C_1, C_3 C_2\} \\ &= \{\mathbf{s}_{c1}, \mathbf{s}_{c2}, \mathbf{s}_{c3}, \mathbf{s}_{c4}, \mathbf{s}_{c5}, \mathbf{s}_{c6}\}.\end{aligned}$$

- Cada ponto (\mathbf{s}_{ci}) tem probabilidade de 1/6 de ser selecionado.
- Supondo, por exemplo, que o ponto $C_1 C_2$ foi selecionado, então

$$S(C_1 C_2) = \{12, 13, 14\} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}.$$

- Condicionado-se a este subespaço ($S(C_1 C_2)$), cada elemento (\mathbf{s}_i) terá probabilidade de 1/3 de ser selecionado aleatoriamente.

Exemplo

- Utilizando-se a probabilidade (marginal) de seleção do ponto ($C_1 C_2$) e condicional de cada par ($\{12,13,14\}$), temos que a probabilidade de sorteio de cada um desses três elementos é

$$P(\mathbf{s}_i) = P(\mathbf{s}_i | \mathbf{s}_{C_i}) P(\mathbf{s}_{C_i}) = \frac{1}{3} \times \frac{1}{6} = \frac{1}{18}.$$

- Ao se considerar todas as possibilidades de pontos de $S_C(\mathcal{U})$ e de cada um dos pontos associados à cada um desses (pares de elementos), teremos bem caracterizado o plano amostral, que pode ser visto em <http://www.ime.unicamp.br/~cnaber/Amostragem%20AC%20em%20dois%20est%C3%A1gios.ods>, lembrando que $\mathbf{d} = (12, 7, 9, 14, 8, 10)^t$ e $\mu = 10$, $\sigma^2 = 34/6$,

Inferência

- Considere o seguinte estimador $\hat{\mu}_{2c} = \frac{1}{a} \sum_{\alpha=1}^a \hat{\mu}_{2\alpha}$, em que a denota o número de conglomeradas selecionados no primeiro estágio e $\hat{\mu}_{2\alpha}$ a média dos elementos selecionados, no segundo estágio, do conglomerado α .
- No exemplo em questão temos que $\hat{\mu}_{2c} = \frac{\hat{\mu}_{s_{i1}} + \hat{\mu}_{s_{i2}}}{2}$, em que $\hat{\mu}_{s_{i1}}$ é a média (no caso o próprio valor) dos elementos do conglomerado $i, i = 1, 2$.
- Perceb que temos duas fontes de aleatoriedade: 1 - seleção de conglomerados, 2 - seleção de elementos dentro dos conglomerados.

Inferência

- Com base nos resultados descritos em

<http://www.ime.unicamp.br/~cnaber/Amostragem%20AC%20em%20dois%20est%C3%A1gios.ods>, temos que a distribuição de probabilidade de $\hat{\mu}_{2c}$ é dada por:

$\hat{\mu}_{2c}$	7,5	8,5	9,5	10	10,5	11	12	13
$P(\hat{\mu}_{2c})$	1/18	1/9	1/6	1/6	1/9	2/9	1/18	1/9

Inferência

- O que nos leva à $\mathcal{E}(\hat{\mu}_{2c}) \approx 10,33$ e $\mathcal{V}(\hat{\mu}_{2c}) = 2$, portanto, $\hat{\mu}_{2c}$ é um estimador viciado. Note que este estimador corresponde a média aritmética (simples) entre as médias dos elementos de cada conglomerado selecionado.

- Considere, agora, o seguinte estimador $\hat{\mu}_{2c1} = \frac{1}{a} \sum_{\alpha=1}^a \frac{B_{\alpha}}{B} \hat{\mu}_{\alpha}$.

Inferência

- Fazendo os desenvolvimentos de modo semelhante ao caso anterior, temos que a distribuição de probabilidade de :

$\hat{\mu}_{2c1}$	7	8	8,25	9,25	9,75	10,25	10,75	11,75	13,5	14,5	15,5
$P(\hat{\mu}_{2c1})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{18}$

Inferência

- O que nos leva à $\mathcal{E}(\hat{\mu}_{2c1}) = 10$ e $\mathcal{V}(\hat{\mu}_{2c1}) \approx 6,92$.
- Isso nos sugere que o estimador é não viciado.
- Mas, $EQM(\hat{\mu}_{2c1}) = \mathcal{V}(\hat{\mu}_{2c1}) = 6,92$ e
 $EQM(\hat{\mu}_{2c}) = \mathcal{V}(\hat{\mu}_{2c}) + B(\hat{\mu}_{2c})^2 \approx 2 + 0,11 = 2,11$, ou seja,
 $EQM(\hat{\mu}_{2c1}) > EQM(\hat{\mu}_{2c})$.

Inferência

- Como a amostragem é feita em dois estágios, primeiro selecionam-se os conglomerados e, depois, os elementos dentro de cada conglomerado, podemos calcular a esperança e a variância usando resultados de distribuições condicionais.
- Por exemplo, para o par $C_1 C_2$, temos que $S_2(C_1 C_2) = \{12, 13, 14\}$, de sorte que: $\tilde{\mu}_{2c1}[12] = 8,25$, $\tilde{\mu}_{2c1}[13] = 9,75$ e $\tilde{\mu}_{2c1}[14] = 13,5$ (valores possíveis do estimador $\hat{\mu}_{2c1}$).

Inferência

- Usando-se o índice 2 para indicar o valor esperado condicional ao particular par de conglomerados, tem-se.

$$\mathcal{E}_2(\hat{\mu}_{2c1}) \equiv \mathcal{E}(\hat{\mu}_{2c1} | C_1 C_2) = \frac{1}{3}(8,25 + 9,75 + 13,5) = \tilde{\mu}_{cd}[C_1 C_2].$$

Inferência

- Estendendo-se os desenvolvimentos para dos demais pares, teremos a seguinte distribuição condicional (note que estamos relaxando a notação apropriada para distribuições de probabilidade e, também, estamos condicionando nos pares de conglomerados $C_i C_j$):

s_{C_i} :	$C_1 C_2$	$C_1 C_3$	$C_2 C_1$	$C_2 C_3$	$C_3 C_1$	$C_3 C_2$
$\mathcal{E}(\hat{\mu}_{cd} C_i C_j)$	10,5	7,5	10,5	12	7,5	12
$P(\mathcal{E}(\hat{\mu}_{cd} = r C_i C_j))$	1/6	1/6	1/6	1/6	1/6	1/6

Inferência

- Usando-se, agora, o índice 1 para indicar a esperança calculada no espaço amostral gerado pelos conglomerados, tem-se

$$\mathcal{E}_1(\hat{\mu}_{2c1}) = \frac{1}{6} (10,5 + 7,5 + 10,5 + 12 + 7,5 + 12) = 10$$

- Esse modus operandi é bastante útil quando se tem vários níveis de seleção no plano amostral.

Notações

- As notações serão basicamente aquelas que foram utilizadas na AC e o planejamento amostral já foi descrito na página 4.
- Lembrando - média populacional: $\mu = \bar{y} = \frac{\tau}{N} = \frac{A\bar{\tau}}{AB} = \frac{\bar{\tau}}{B}$
- Estimador (utilizado) $\hat{\mu}_{2c1} = \frac{1}{a} \sum_{\alpha=1}^a \frac{B_{\alpha}}{B} \hat{\mu}_{\alpha}$, em que $\hat{\mu}_{\alpha}$ é a média dos elementos selecionados do conglomerado α (que também fora selecionado).
- Para outros estimadores veja as referências, incluindo o livro texto.

Inferência

- Valor esperado e variância.
- Sejam X e Y duas variáveis aleatórias definidas no mesmo espaço de probabilidade. Então

$$\mathcal{E}_X(X) = \mathcal{E}_Y(\mathcal{E}_{X|Y}(X|Y))$$

e

$$\mathcal{V}(X) = \mathcal{E}_Y(\mathcal{V}_{X|Y}(X|Y)) + \mathcal{V}_Y(\mathcal{E}_{X|Y}(X|Y)).$$

Inferência

- Vamos utilizar as notações definidas anteriormente, ou seja, $\mathcal{E}((.)|S_C(\mathcal{U})) \equiv \mathcal{E}_2((.))$ (entre os elementos, condicionado no espaço amostral gerado pelos conglomerados) e $\mathcal{E}_1(.)$ (esperança em função do espaço amostral gerado pelos conglomerados).
- Assim, temos que

$$\mathcal{E}(\hat{\mu}_{2c1}) = \mathcal{E}_1(\mathcal{E}_2(\hat{\mu}_{2c1})). \quad (1)$$

Inferência

- Mas, dentro do α -ésimo conglomerado (sorteado com B_α unidades), os b_α elementos são selecionados segundo uma AAS_c , assim, de http://www.ime.unicamp.br/~cnaber/aula_AAS%20com%20reposicao%20parte%201%20Amost%20S%202018.pdf

$$\mathcal{E}_2(\hat{\mu}_{2c1}) = \frac{1}{a} \sum_{\alpha=1}^a \frac{B_\alpha}{B} \mathcal{E}_2(\hat{\mu}_\alpha) = \frac{1}{a} \sum_{\alpha=1}^a \frac{B_\alpha}{B} \mu_\alpha \quad (2)$$

- Atenção: Note, no entanto, que μ_α é uma variável aleatória que representa a média do α -ésimo conglomerado sorteado (pois serão sorteados $b_\alpha \leq B_\alpha$ elemento daquele).

Inferência

- Suponha uma população formada de A elementos (conglomerados) e defina $x_\alpha = \frac{B_\alpha}{B} \mu_\alpha$, $\alpha = 1, 2, \dots, A$.
- Assim, a média, sob uma amostra de a unidades (AAS_c), digamos,

$$\bar{X} = \frac{1}{a} \sum_{\alpha=1}^a X_\alpha = \frac{1}{a} \sum_{\alpha=1}^a \frac{B_\alpha}{B} \mu_\alpha \text{ é tal que}$$

$$\mathcal{E}_1(\bar{X}) = \frac{1}{A} \sum_{\alpha=1}^A \frac{B_\alpha}{B} \mu_\alpha = \mu \quad (3)$$

(conforme visto em http://www.ime.unicamp.br/~cnaber/aula_AC_Amost_2S_2018.pdf).

Inferência

- Logo, de (2) e (3) em (1), vem que:

$$\mathcal{E}(\hat{\mu}_{2c1}) = \mathcal{E}_1(\bar{X}) = \mathcal{E}_1\left(\frac{1}{a} \sum_{\alpha=1}^a \frac{B_{\alpha}}{B} \mu_{\alpha}\right) = \mu.$$

- Portanto, $\hat{\mu}_{2c1}$ é um estimador não viciado para μ .

Inferência

- Para a variância, temos que:

$$\mathcal{V}(\hat{\mu}_{2c1}) = \mathcal{E}_1(\mathcal{V}_2(\hat{\mu}_{2c1})) + \mathcal{V}_1(\mathcal{E}_2(\hat{\mu}_{2c1})) \quad (4)$$

- Primeiramente, note, pelos mesmos argumentos anteriores, que

$$\begin{aligned} \mathcal{V}_1(\bar{X}) &= \mathcal{V}_1\left(\frac{1}{a} \sum_{\alpha=1}^a \frac{B_{\alpha}}{B} \mu_{\alpha}\right) = \frac{\sigma_x^2}{a} = \frac{1}{aA} \sum_{\alpha=1}^A \left(\frac{B_{\alpha}}{B} \mu_{\alpha} - \mu\right)^2 \\ &= \frac{\sigma_{ect}^2}{a} \end{aligned} \quad (5)$$

Inferência

- Por outro lado, pelo mesmos argumentos que nos levaram à (2) (sobs AAS_c e lembrando que tomaremos uma amostra de b_α elementos do conglomerado α), temos que:

$$\begin{aligned} \mathcal{V}_2(\widehat{\mu}_{2c1}) &= \frac{1}{a^2} \sum_{\alpha=1}^a \left(\frac{B_\alpha}{B} \right)^2 \mathcal{V}_2(\widehat{\mu}_\alpha) \\ &= \frac{1}{a^2} \sum_{\alpha=1}^a \left(\frac{B_\alpha}{B} \right)^2 \frac{\sigma_\alpha^2}{b_\alpha} = \frac{1}{a} \sum_{\alpha=1}^a v_\alpha = \bar{V} \end{aligned}$$

em que $v_\alpha = \frac{1}{a} \left(\frac{B_\alpha}{B} \right)^2 \frac{\sigma_\alpha^2}{b_\alpha}$ é uma variável aleatória auxiliar (semelhante à X_α).

Inferência

- Com efeito, utilizando raciocínio análogo àquele utilizado para X_α , vem que:

$$\begin{aligned} \mathcal{E}_1(\mathcal{V}_2(\widehat{\mu}_{2c1})) &= \mathcal{E}_1\left(\frac{1}{a^2} \sum_{\alpha=1}^a \left(\frac{B_\alpha}{\overline{B}}\right)^2 \frac{\sigma_\alpha^2}{b_\alpha}\right) = \mathcal{E}_1(\overline{V}) = \overline{v} = \frac{1}{A} \sum_{\alpha=1}^A v_\alpha \\ &= \frac{1}{aA} \sum_{\alpha=1}^A \left(\frac{B_\alpha}{\overline{B}}\right)^2 \frac{\sigma_\alpha^2}{b_\alpha} \end{aligned} \quad (6)$$

em que $\overline{V} = \frac{1}{a} \sum_{\alpha=1}^a V_\alpha$ é a média de uma amostra de tamanho a de uma população com os valores V_1, \dots, V_A .

Inferência

- Para a segunda parte, temos de (2) e (5), vem que:

$$\begin{aligned} \mathcal{V}_1(\mathcal{E}_2(\hat{\mu}_{2c1})) &= \mathcal{V}_1\left(\frac{1}{a} \sum_{\alpha=1}^a \frac{B_\alpha}{B} \mu_\alpha\right) = \mathcal{V}_1(\bar{X}) = \frac{\sigma_x^2}{n} \\ &= \frac{1}{aA} \sum_{\alpha=1}^A \left(\frac{B_\alpha}{B} \mu_\alpha - \mu\right)^2 \end{aligned} \quad (7)$$

Inferência

- De (6) e (7), em (4), finalmente, temos que:

$$\mathcal{V}(\hat{\mu}_{2c1}) = \frac{1}{aA} \sum_{\alpha=1}^A \left(\frac{B_{\alpha}}{B} \mu_{\alpha} - \mu \right)^2 + \frac{1}{aA} \sum_{\alpha=1}^A \left(\frac{B_{\alpha}}{B} \right)^2 \frac{\sigma_{\alpha}^2}{b_{\alpha}}$$

ou seja, a variância do estimador é a soma da variância entre e a variância dentro dos conglomerados, diferentemente do mesmo estimador sob AC (um único estágio) que dependia apenas da variância entre conglomerados.

Inferência

- Podemos escrever a variância acima como

$$\mathcal{V}(\hat{\mu}_{2c1}) = \frac{\sigma_{ect}^2}{a} + \frac{\sigma_{2dc}^2}{a\psi}$$

em que $\sigma_{2dc}^2 = \frac{1}{A} \sum_{\alpha=1}^A \left(\frac{B_{\alpha}}{B}\right)^2 \frac{\psi}{b_{\alpha}} \sigma_{\alpha}^2$ e $\psi = \frac{1}{A} \sum_{\alpha=1}^A b_{\alpha}$.

- Estimadores não viciados, para as duas variâncias acima, são dados, respectivamente, por

- $\hat{\sigma}_{2dc}^2 = \frac{1}{a} \sum_{\alpha=1}^a \left(\frac{B_{\alpha}}{B}\right)^2 \frac{\psi}{b_{\alpha}} \hat{\sigma}_{\alpha}^2$, em que $\hat{\sigma}_{\alpha}^2$ é a variância amostral do conglomerado $\alpha = 1, 2, \dots, A$, dividida por $b_{\alpha} - 1$.

- $\hat{\sigma}_{ect}^2 = \hat{\sigma}_{2ect}^2 - \frac{\hat{\sigma}_{2dc}^2}{\psi}$, em que $\hat{\sigma}_{2ect}^2 = \frac{1}{a-1} \sum_{\alpha=1}^a \left(\frac{B_{\alpha}}{B} \hat{\mu}_{\alpha} - \hat{\mu}_{2c1}\right)^2$

Intervalos de confiança, testes de hipótese e tamanho amostral

- Assim como no caso de AC, os resultados assintóticos usuais são válidos quando a e A são suficientemente grandes, mas também b_α , $\alpha = 1, 2, \dots, A$.
- Tais resultados podem ser utilizados para constuir IC's, testar hipóteses e determinar tamanhos de amostra de modo semelhante ao que foi feito para AC, veja http://www.ime.unicamp.br/~cnaber/aula_AC_P2_Amost_2S_2018.pdf.

Coeficiente de correlação intraclasse e EPA

- Pesquisa sobre os dois outros estimadores e como os três se comportam sob conglomerados de igual tamanho (populacional e amostralmente).
- Sob conglomerados de mesmo tamanho ($B_\alpha = B$) (populacional) e ($b_\alpha = b$) (sorteado), $\forall \alpha$, temos que

$$(AC2E)EPA(\hat{\mu}_{2c}) = 1 + \rho_{int}(b - 1)$$

$$(AC)EPA(\hat{\mu}_c) = 1 + \rho_{int}(B - 1)$$

Coefficiente de correlação intraclasse e EPA

- Em que $\hat{\mu}_{2c}$ e $\hat{\mu}_c$ são, respectivamente, o estimador (aos quais se igualam os três) sob os planos AC2E e AC, quando todos os conglomerados (populacionais e amostrais) possuem o mesmo tamanho.
- Note que, se $\rho_{int} > 0$, então:

$$\frac{\mathcal{V}_{AC2E_1}(\hat{\mu}_{2c})}{\mathcal{V}_{AC_1}(\hat{\mu}_c)} = \frac{1 + \rho_{int}(b - 1)}{1 + \rho_{int}(B - 1)} \leq 1$$